UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Single-cell transcriptomics in unravelling the molecular complexity of immunity in human disease

Ana Marta Fernandes Bica

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Prof.ª Dr.ª Maria Fernanda Nunes Diamantino
Doutor Nuno Luís Barbosa Morais

2019

Acknowledgments

Firstly, I would like to express my sincere gratitude to my advisor Nuno Morais for the continuous support of my work, for his guidance, motivation, and immense knowledge. I could not have imagined having a better advisor and mentor.

I would like to thank Karine Serre, without your precious support it would not have been possible to conduct this project. All the times spent in your team meeting and all the things you taught me on cancer immunology inspired me to begin the next chapter in my life. I couldn't have done it without you.

My sincere thanks also goes to professor Fernanda Diamantino, for enlightening me in biostatistics concepts that were fundamental for this project, and will continue to be throughout my career as a computational biologist. Your positivity and encouragement brightened my days.

A very special gratitude goes out to all down at the Disease Transcriptomics lab. Marie, Mariana, Nuno, Sofia, Sara and Arthur, you are the best colleagues and friends anyone could ask for, thank you.

I would also like to thank the people that supported me continuously throughout my life, specially my mother, who passed on the science curiosity genes to me, and Fábio, for always supporting me in my aspirations.

Finally, a big thank you to my friends. Micaela, one of the best parts of this master's was getting to know you. There were some tough semesters, but you helped me through them. Jorge, my academic accomplice, you are an amazing friend, and I hope we'll keep on sharing academic insights for many years to come.

Resumo

O sistema imunitário engloba milhões de células que formam uma estrutura dinâmica e em comunicação, com o objetivo de defender o hospedeiro contra a entrada de agentes patogénicos e outras ameaças, como o aparecimento de células cancerígenas. No ser humano, a resposta imunitária envolve diversos tipos de células e, para cada tipo, estados celulares diferentes, interagindo entre si, de forma a manter e a proteger a função e integridade do organismo.

Ao longo da vida o sistema imunitário vigia continuamente o organismo, através de um equilíbrio que envolve células imunitárias efectoras e reguladoras. No entanto, com o envelhecimento, ocorre um declínio gradual deste sistema, definido como imunosenescência. Esta deterioração leva a alterações nas proporções de diferentes tipos de células imunitárias no organismo e nas suas competências, o que, por sua vez, contribui para o aumento da prevalência de cancro, bem como para a propensão para um estado de inflamação crónica de baixo grau, implicado noutras doenças relacionadas com o envelhecimento, tais como as doenças neurodegenerativas. De facto, a idade é o principal fator de risco para o desenvolvimento da doença de Alzheimer e da doença de Parkinson. A sobreposição temporal entre o envelhecimento e a neurodegeneração tem alimentado um debate contínuo sobre se todos nós somos suscetíveis a desenvolver uma doença neurodegenerativa se vivermos tempo suficiente. Esta hipótese é sustentada pelo facto de o cérebro envelhecido apresentar várias lesões que não estão presentes no cérebro de pessoas mais jovens e de essas lesões se assemelharem a uma versão de grau inferior às encontradas nas doenças neurodegenerativas mais comuns. Muitos dos mecanismos implicados nas doenças neurodegenerativas são paralelos às mudanças que ocorrem com o envelhecimento e a maioria dos cérebros de idosos apresenta alterações específicas que podem ser ligadas a um certo nível de neurodegeneração, tal como a agregação de placas proteicas tóxicas e a neuro-inflamação.

A neuro-inflamação consiste numa desregulação do sistema imunitário, associada à ativação e hiper-reactividade da microglia, o principal tipo de células imunitárias no sistema nervoso central. Os astrócitos são outro tipo de célula não-neuronal presente no sistema nervoso central, que participa na constituição da barreira hematoencefálica e providencia suporte aos neurónios, entre outras funções. Estas células também se tornam excessivamente reativas com o envelhecimento. A neuro-inflamação é, assim, uma potencial causa das alterações funcionais que ocorrem durante o envelhecimento normal e patológico, tendo um efeito tremendo no aumento da suscetibilidade às doenças neurodegenerativas. A exacerbação, com a idade, de processos inflamatórios no sistema nervoso central leva à perda da homeostase e, consequentemente, à disfunção ou morte de células neuronais, tal como observado na doença de Alzheimer e na doença de Parkinson. Contudo, uma minoria de pessoas ultrapassa a idade dos 80 anos sem mostrar sinais de debilitação cognitiva. Estas pessoas são a prova da existência de mecanismos compensatórios, que lhes permitem envelhecer saudavelmente e manter uma cognição normal.

O envelhecimento é também um dos principais factores de risco do cancro. O cancro da mama é uma doença particularmente associada às mulheres mais velhas, raramente ocorrendo antes dos 30 anos de idade, e com maior prevalência acima dos 60 anos de idade. Todos os anos, mais de 1,5 milhões de mulheres são diagnosticadas com cancro da mama, tornando-o o cancro mais comum, bem como o segundo mais mortal, nas mulheres, a nível mundial.

As terapias tradicionais do cancro da mama incluem quimioterapia e radioterapia. Contudo, o tratamento desta doença continua a ser um desafio devido à sua natureza heterogénea. Aliás, existem vários subtipos de cancro da mama, sendo o mais agressivo o cancro da mama triplo-negativo. Ainda não existem medicamentos e/ou terapias especificamente direcionadas ao tratamento deste tipo de cancro, o que

enfatiza a necessidade de explorar terapias alternativas, tal como a imunoterapia, tendo em conta que a importância do sistema imunitário no cancro da mama é inequívoca.

O microambiente tumoral é composto por matriz extracelular e células infiltradas na massa tumoral. Estas incluem uma proporção elevada de células tanto do o sistema imunitário inato como do adaptativo. As células imunitárias no microambiente tumoral foram já descritas como sendo determinantes na iniciação, progressão e metastização do cancro. Apesar da existência de subtipos diferentes de cancro da mama, e da variabilidade interindividual de doentes, estudos recentes, focados no microambiente tumoral, demonstraram a existência de padrões de infiltração de células imunitárias correlacionados com prognóstico (negativo ou favorável) em doentes com cancro da mama. Por exemplo, existe uma associação entre a infiltração de linfócitos, nomeadamente células T CD8+, e prognóstico favorável da doente, dado que estas células possuem funções citotóxicas capazes de eliminar células tumorais. Desta forma, as células T, no microambiente tumoral, já foram extensivamente estudadas e relacionadas como um componente fundamental do mesmo, devido a ensaios clínicos recentes que demonstraram a capacidade de controlar a progressão do cancro nalguns doentes ao manipular estas células.

Contudo, as células mieloides, que englobam um conjunto de células do sistema imunitário inato, permanecem menos estudadas do que as células T, apesar de constituírem uma proporção significativa das células imunitárias infiltradas no microambiente tumoral. Dessas células, destacam-se os macrófagos, que podem atingir proporções superiores a 50% da própria massa tumoral na mama. Estas células apresentam polarizações diferentes conforme os estímulos do microambiente, podendo adotar um fenótipo que inibe o crescimento do tumor (anti-tumoral) ou que o favorece (pro-tumoral). A presença de macrófagos pró-tumorais no microambiente tumoral tem vindo a ser descrita como um indicador de prognóstico negativo no cancro, incluindo no cancro da mama. Por outro lado, os macrófagos anti-tumorais têm a capacidade de induzir a regressão dos tumores, pelo que, geralmente, encontram-se associados a um prognóstico favorável no cancro da mama. A existência de macrófagos com funções antagónicas torna-os num tópico de investigação ativo na área do cancro da mama, com o objectivo de desenvolver novas imunoterapias, direcionadas à diminuição dos números de macrófagos pro-tumorais, ou à re-polarização dos mesmos em macrófagos anti-tumorais.

Assim, é importante caracterizar a heterogeneidade celular do sistema imunitário, possibilitando o conhecimento dos processos biológicos fundamentais que ocorrem no envelhecimento saudável e na doença. Para tal, é possível utilizar *single-cell RNA sequencing*, uma abordagem que utiliza tecnologias de sequenciação do transcritoma de uma única célula, permitindo medir a distribuição dos níveis de expressão de cada gene numa população de células individuais e melhor entender os padrões de expressão génica em tecidos complexos, como o cérebro e o microambiente tumoral.

O desenvolvimento de protocolos de s*ingle-cell RNA sequencing* foi motivado pela necessidade de estudar condições em que apenas uma pequena quantidade de material se encontrava disponível, tal como o desenvolvimento embrionário. Contudo, o aperfeiçoamento de protocolos e das plataformas de sequenciação permitiu um aumento no número de células utilizadas nestes ensaios, podendo chegar às centenas de milhares de células por estudo. Esta evolução contínua tem vindo a melhorar radicalmente a dissecção da heterogeneidade de populações celulares, particularmente na estimação da infiltração de células imunitárias em tumores sólidos, e na área das neurociências, permitindo caracterizar a grande diversidade de células neuronais e não-neuronais em várias regiões do sistema nervoso central.

Este trabalho consistiu em duas partes, ambas envolvendo a análise de dados públicos de s*ingle-cell RNA sequencing*. Na primeira parte, focámo-nos em desenvolver uma *pipeline* de análise computacional, de forma a estudar as abundâncias relativas de células neuronais e não-neuronais no

cérebro, e como a abundância destas populações se correlaciona com o envelhecimento e patologias neurológicas.

Com esta análise demonstrámos que, para alguns tipos específicos de tecido cerebral, ocorre um decréscimo na proporção de neurónios com a idade, concomitante com um aumento na proporção de astrócitos. Esta alteração de proporções é acentuada na doença de Alzheimer e na doença de Parkinson, nas quais os tecidos afetados pela neurodegeneração demonstram uma proporção relativa inferior de neurónios, bem como uma proporção relativa superior de astrócitos, quando comparados com tecidos cerebrais sem doença.

Também demonstrámos que o sistema nervoso central não é um local totalmente imuno-privilegiado pois conseguimos estimar as abundâncias absolutas de diferentes tipos de células imunitárias no cérebro. Aliás, esta análise revelou que, as células T CD4+ de memória são as células imunitárias que se infiltram em maior quantidade no cérebro, existindo uma variabilidade elevada entre indivíduos.

A segunda parte deste trabalho teve como objectivo avaliar a diversidade celular e a assinatura molecular de macrófagos infiltrados no microambiente tumoral no cancro da mama, bem como caracterizar os padrões de infiltração de células imunitárias na massa tumoral e o modo como estes se encontram associados com a idade e o prognóstico.

Com esta análise, identificámos grupos de macrófagos infiltrados na massa tumoral com fenótipos diferentes, tais como macrófagos com funções pró-inflamatórias anti-tumorais e macrófagos alternativamente ativados, ou seja, com funções anti-inflamatórias pró-tumorais. Por outro lado, descobrimos ainda um grupo de macrófagos transcricionalmente ativos, que não se assemelham a nenhuma das polarizações previamente descritas, tratando-se, possivelmente, de um novo estado por descrever.

Verificámos ainda a ocorrência de um aumento significativo com a idade na proporção relativa de macrófagos pró-tumorais no microambiente tumoral do cancro da mama. Estes macrófagos têm a capacidade de suprimir a resposta anti-tumoral das células T CD8+ citotóxicas. Ao comparar grupos de tumores com uma proporção relativa elevada/baixa de macrófagos pró-tumorais, e grupos de tumores com uma proporção relativa baixa/elevada de células T CD8+, verificámos que a infiltração de macrófagos pró-tumorais parece estar associada com o processo biológico de transição epitelial-mesenquimal, o qual está envolvido no potencial metastático dos tumores malignos. Em contraste, a infiltração de células T CD8+ encontra-se associada ao reconhecimento do tumor e à consequente elicitação de uma resposta imunitária ativa.

Palavras-chave: *single-cell RNA sequencing*, envelhecimento, imunosenescência

Abstract

Throughout the course of life, the immune system keeps surveilling the organism for foreign pathogens and cancerous cells. However, with ageing there is a gradual decline of the immune system fitness, which is defined as immunosenescence. This deterioration leads to alterations in the proportions of different immune cell types in the organism and in their capabilities. Moreover, it contributes to the increased prevalence of cancer as well as a propensity of a chronic low-grade inflammatory state implicated in other age-related diseases, such as neurodegenerative ones.

The immune system comprises a multitude of different cell types and states. It is important to assess this cell heterogeneity in order to understand fundamental biological processes in healthy ageing and disease. One way to do this is through single-cell RNA sequencing, an approach that uses sequencing technologies to profile the transcriptome of a single cell, thereby allowing to measure the distribution of expression levels for each gene across a population of individual cells, and to better understand gene expression patterns in complex heterogeneous tissues, such as the brain and the tumour microenvironment.

Using publicly available single-cell RNA sequencing datasets, the first part of this work was focused on developing a computational analysis pipeline to study the relative abundance of neuronal and non-neuronal cells in the brain and how they correlate with ageing and neurological health.

We found that, for some brain tissues, there is a decrease in the proportion of neurons with ageing, concomitant with an increase in the proportion of astrocytes. This shift in proportions is accentuated in Alzheimer's disease and Parkinson's disease, in which specific brain tissues affected by neurodegeneration show a relatively lower proportion of neurons and a relatively higher proportion of astrocytes, when compared with controls. We also demonstrated that the central nervous system in not totally an immune-privileged tissue without infiltration of blood-leucocytes, by estimating absolute abundances of different immune cell types in the brain. This revealed that resting CD4+ memory T cells present the highest proportion of brain-infiltrating immune cells, with a relatively high level of variability between individuals.

In the second part of this work, our goal was to evaluate the cellular diversity and molecular signature of breast tumour-associated macrophages, and to understand how the intra-tumoural diversity and functionality of infiltrating immune cell types was associated with age and prognosis.

By implementing single-cell RNA sequencing data analysis tools, we were able to discern groups of tumour-infiltrating macrophages with different phenotypes, such as the classically activated polarization (anti-tumour) macrophages and the alternatively activated polarization (pro-tumour) ones. On the other hand, we found a group of transcriptionally activate macrophages that do not resemble any of the previously described polarizations, possibly being a new unstudied state. We also found that there is a significant increase in the relative proportion of breast tumour-infiltrating pro-tumour macrophages with ageing. These macrophages are known to suppress CD8+ cytotoxic T cell-mediated anti-tumour immune responses. By comparing groups of breast tumour bulk RNA-sequencing samples with a high/low proportion of pro-tumour macrophages, and a low/high proportion of CD8+ T cells, we found that infiltration of pro-tumour macrophages is associated with the cancer metastasis hallmark (epithelial-mesenchymal transition). In contrast the tumours with higher infiltration of CD8+T cells, were associated with the recognition of the breast tumour as non-self and the consequent elicitation of an active immune response.

Keywords: single-cell RNA sequencing, ageing, immunosenescence

Table of contents

List of figures

List of tables

Abbreviation List

Aβ - β-amyloid peptide

AD - Alzheimer's disease

APCs - Antigen-presenting cells

APP - Amyloid precursor protein

BBB - Blood-brain barrier

CD - Cluster of differentiation

CNS - Central nervous system

DC - Dendritic cell

ER - Oestrogen receptor

FDCSP - Follicular dendritic cell secreted protein

GFAP - Glial fibrillary acidic protein

GM-CSF - granulocyte-macrophage colony-stimulating factor

GO – Gene ontology

HER2 - Human Epidermal growth factor Receptor 2

HDG - High dropout genes

HVG - Highly variable genes

IHC - Immunohistochemistry

LCM - Laser capture microdissection

M-CSF - Monocyte colony stimulating factor

M0 - Non-polarized macrophages

MDSC - Myeloid-derived suppressor cell

NK - Natural killer

PCA - Principal component analysis

PD - Parkinson's disease

pDC - Plasmacytoid dendritic cell

PR - Progesterone receptor

TAM - Tumour-infiltrating macrophage

Th - Helper T

TIL - Tumour-infiltrating lymphocyte

TME - Tumour microenvironment

TNBC - Triple negative breast cancer

Treg - Regulatory T

t-SNE - t-distributed stochastic neighbour embedding

1. The Immune System

The immune system consists of millions of cells that form a dynamic communicating structure, with the goal of defending the host against infecting pathogens and other agents, such as malignant transformed cells. The immune response in human health and disease involves a multitude of different cell types and states, interacting amongst each other and with non-immune cells, to maintain and protect tissue function and integrity. The distinct elicited immune responses, along with the immune cells that are part of this complex network, can be classified into two major components: the innate immune system and the adaptive immune system (Goldman and Prabhakar 1996).

1.1. Innate Immune system

The innate immune response is an immediate and non-specific defence mechanism. Besides including anatomical barriers that avoid the entrance of infectious agents in the organism, this response is dependent on the recruitment of natural killer and myeloid cells. Natural killer (NK) cells are effector lymphocytes with cytotoxicity and cytokine-producing effector functions, directed at killing infected host cells and tumour cells, limiting their spread (Vivier et al. 2008). Myeloid cells can be phagocytic (neutrophils, monocytes, dendritic cells and macrophages) or inflammatory mediator-releasing cells (basophils, mast cells, and eosinophils) (Delves and Roitt 2000).

Monocytes are blood-circulating phagocytes that can develop into macrophages or dendritic cells after migrating into tissues (Karlmark, Tacke and Dunay 2012).

Dendritic cells (DC) are potent antigen-presenting cells (APCs), acting in tissues that are in contact with the external environment, linking the innate and the adaptive arms of the immune system (Mellman 2013). Myeloid-derived DCs are usually considered a distinct cell type from plasmacytoid dendritic cells (pDC), given that they may not descend from the myeloid lineage (Reizis 2019).

Macrophages are highly adherent and motile phagocytes that patrol tissues for potential pathogens. They also produce large amounts of pro-inflammatory cytokines and chemokines, and are able to recruit lymphocytes through antigen-presentation, assisting the initiation of the adaptive immune response (Goldman and Prabhakar 1996).

Monocytes and macrophages from different tissues, together with their precursors, constitute the mononuclear phagocyte system, in which monocytes replace resident macrophages in all major organs, by adopting specific gene expression profiles, which translate into distinct functions. Kupffer cells in the liver and alveolar macrophages in the lung are two examples of specialized monocyte-derived macrophages (Hume, Irvine and Pridans 2019).

Neutrophils, basophils and eosinophils are granulocytic cells, i.e., they contain cytoplasmic granules filled with antimicrobial products, inflammation mediators, and cytotoxic proteins, respectively. While neutrophils engulf and digest pathogens, basophils and eosinophils degranulate the content of their granules in response to parasitic infections and allergies. Mast cells have similar functional characteristics as basophils. However, unlike the latter, they reside in tissue instead of the bloodstream (Goldman and Prabhakar 1996).

In the central nervous system (CNS), the major resident myeloid cells are microglia. Microglia survey the microenvironment and release trophic factors which are important for neuronal cell survival. Their

phagocytosis capability is also important in synaptic homeostasis and clearance of cellular debris resulting from injury. Although functionally similar to macrophages, microglia originate from the yolk sac and populate the CNS prior to its vasculogenesis (Nayak, Roth and McGavern 2014). Microglia, together with oligodendrocytes and astrocytes, are the major components of glia, i.e., the main non-neuronal cells of the CNS, that are key in maintaining homeostasis and neuronal function. Representing the larger fraction of glia in the brain, astrocytes are responsible for the maintenance of ion homeostasis and the blood-brain barrier (BBB), production of neurotrophic factors, participation in the formation, maturation and elimination of synapses, and uptake of neurotransmitters. Furthermore, astrocytes play a role in local immune regulation, by releasing inflammatory mediators that activate and amplify the initial innate immune response, and by altering BBB permeability, allowing the entrance of peripheral blood immune cells in the brain parenchyma (Vasile, Dossi and Rouach 2017; Farina, Aloisi and Meinl 2007).

1.2. Adaptive immune system

The adaptive immune response, also known as the acquired immune response, is mediated by specialized cells, capable of targeting pathogens more accurately than the innate system, and of a long-term response, enabled by the generation of immunological memory. This leads to a more efficient immune response when the pathogen is reencountered (Delves and Roitt 2000). These specialized cells are lymphocytes, namely antigen-specific B and T cells.

Activated B cells (short-lived plasma cells) secrete neutralizing immunoglobins that bind to and lead to the elimination of pathogens. Like macrophages and dendritic cells, they are also APCs. Some B cells become pathogen-experienced memory B cells, which are important in eliciting an enhanced immune response in the case of re-infection (Goldman and and Prabhakar 1996; Kurosaki, Kometani and Ise 2015).

T cells can be categorized into three distinct subsets: effector T cells, regulatory T cells and memory T cells. Effector T cells actively respond to stimuli and can be further divided into helper T cells and cytotoxic T cells. Helper T cells express the CD4 molecule and regulate the immune response, by producing cytokines that stimulate phagocytic cells, aid B cell proliferation and differentiation, and assist other effector T cells in cell-mediated immunity. Cytotoxic T cells express the CD8 molecule and directly kill infected or altered cells, through the production of cytolytic enzymes. Regulatory T cells, also known as suppressor T cells, express CD4 and are involved in preventing immune overreaction, by supressing T and B cell activity (Cano and Lopera 2013). Memory T cells can be either CD4+ or CD8+, and represent long-lived populations of T cells that can rapidly differentiate into effector T cells, when re-encountering pathogens to which they were previously exposed (Omilusik and Goldrath 2017).

In summary, the innate response is a nonspecific response that is not altered by the number of times the same pathogen is encountered. The adaptive response is a highly specific and long-lasting response that is enhanced when the pathogen is re-encountered. Both systems are connected in a complex net of relationships that allows them to work together to neutralize potential threats to the organism. Thus, the failure in any component of the immune system may lead to a deficient immunosurveillance state.

2. Neurodegenerative diseases

For the first time in history, we are rapidly approaching a shift in the world's population age distributions. It is expected that, by the year 2030, people older than 60 years will have already outnumbered children under the age of 10, and, by 2050, they will also outnumber children and young adults between 10 to 24 years of age (United Nations 2017).

Although the desired increase in life expectancy is a product of the success of medical advancements, public health and socio-economical improvements, global ageing is expected to lead to an accelerated increase in the prevalence of neurodegenerative disorders, such as Alzheimer's disease (AD) and Parkinson's disease (PD) (Erkkinen, Kim and Geschwind 2017). Both these diseases are sources of growing morbidity and mortality rates, particularly in the elderly.

AD and PD present different clinical symptomatology and progression, albeit there are overlapping features in their pathological processes, such as defects in proteostasis and immune regulation, concomitant with a common inflammatory mechanism, implicated in the chronic progression of most neurodegenerative diseases (Gao and Hong 2008).

2.1. Alzheimer's disease

Late-onset AD is the most common cause of dementia in the world. It is characterised by progressive short-term memory loss, followed by language, visuospatial and executive function impairment. Patients in later stages may also display motor dysfunction. The disorder eventually culminates in death, with patients living, on average, 8 to 10 years, after diagnosis (Schachter and Davis 2000).

The pathology of AD is associated with the accumulation of neurofibrillary tangles and deposition of neuritic plaques in the brain (Figure 2.1B) (Zheng and Koo 2006). Neuritic plaques consist of insoluble deposits of β-amyloid peptide (Aβ), a fragment of the larger amyloid precursor protein (APP). Currently, it is still unclear whether neuritic plaques are a cause or a consequence of AD. Previous studies report contrasting hypotheses, for example, either stating that dysfunction in the metabolism of APP, with subsequent increase in the insoluble Aβ, is responsible for AD (Schachter and Davis 2000), or that plaques exist in both the cognitive impaired elderly and elderly which show no loss of cognitive ability (Makin 2018).

Neurofibrillary tangles are intracellular accumulations of modified tau protein, secondary to Aβ deposition. In the healthy brain, tau proteins are microtubule stabilizers. However, in AD brain, tau proteins are hyperphosphorylated, leading to their self-polymerization and to the formation of tangles (Kolarova et al. 2012).

In spite of their unclear role in the pathology of AD, it is known that neurofibrillary tangles and neuritic plaques are neurotoxic and can lead to neuron degeneration, either by directly disrupting signalling and obstructing cell function, or by indirectly causing inflammation and oxidative stress (Farias et al. 2011; Yankner, Duffy and Kirschner 1990; Huang, Zhang and Chen 2016).

The distribution pattern of neurofibrillary tangles and neuronal alterations in a subject's brain is used to verify the stage of AD progression (Braak and Braak 1991). In Braak stages I and II, neuronal alterations are confined to the transentorhinal region of the brain (Figure 2.1A). Stages III and IV occur when there

is involvement of the transentorhinal region and the hippocampus, and V and VI when there is extensive neocortical degeneration (Dossi, Vasile and Rouach 2018).

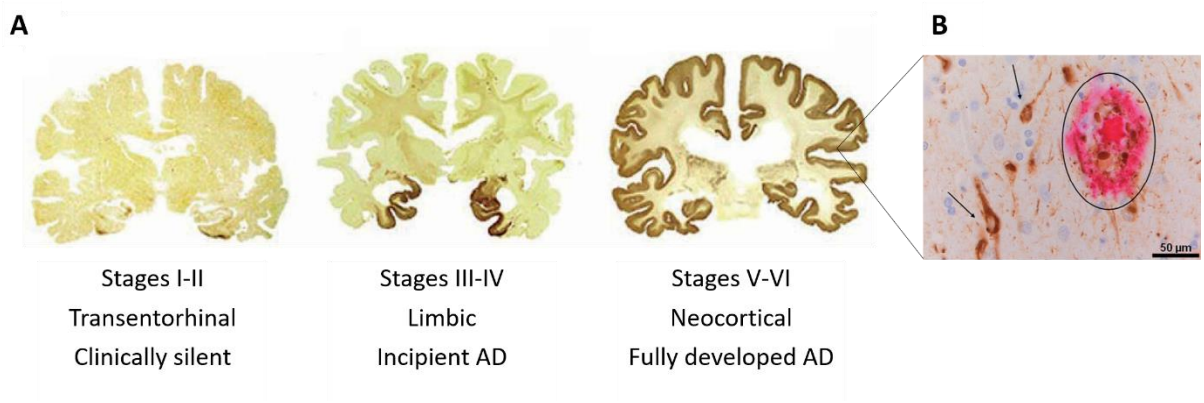

Figure 2.1: Disease progression in Alzheimer's disease. ***A*** *- Distribution pattern of neurofibrillary tangles and abnormal neurites;* ***B*** *- Immunohistochemical staining of intracellular neurofibrillary tangles (brown, indicated by arrows) and extracellular neuritic plaques (pink, indicated with a circle). Scale bar 50 µm. Adapted from Vies 2016.*

Currently, there is no treatment to stop AD progression, with existing pharmaceutical therapies, e.g. acetylcholinesterase inhibitors and memantine, only providing modest cognitive and functional benefits (Pierce, Bullain and Kawas 2017).

2.2. Parkinson's disease

PD is the second most common neurodegenerative disease. PD mainly affects the motor system and is characterised by slow movement and an impaired ability to move the body swiftly on command. Manifestations such as tremors and muscle stiffness are also common in PD. Dementia can also be present in later stages (Costantini, D'Angelo and Reale 2018). The median age of onset is 60 years, with patients living, on average, 15 years after diagnosis (Erkkinen, Kim and Geschwind 2017). However, the age of the patient seems to be responsible for the progression of clinical symptoms, rather than the age of disease onset (Hindle 2010).

Although the cause of PD is still unknown, motor symptoms are attributed to the loss of dopaminergic neurons in the *substantia nigra pars compacta*, resulting in a decrease of dopamine levels in the brain (Costantini, D'Angelo and Reale 2018). Another characteristic feature of PD is the presence of Lewy bodies, resulting from accumulation of α-synuclein and ubiquitin aggregates in the neuronal cytoplasm. α-synuclein is also present in neuronal processes, as well as in astrocytes and oligodendrocytes (Shults 2006).

Braak staging in PD is performed according to the degree and localization of Lewy body accumulation, neurodegeneration and consequent clinical symptomology. In stage I, pathological changes occur only in the dorsal motor nucleus and olfactory bulb. Stage II involves Lewy body formation in the pons and medulla. In stage 3 and 4, patients exhibit clinical motor symptoms, and stages 5 and 6 involve degeneration of the neocortex, leading to cognitive impairment and dementia (Figure 2.2) (Hindle 2010).

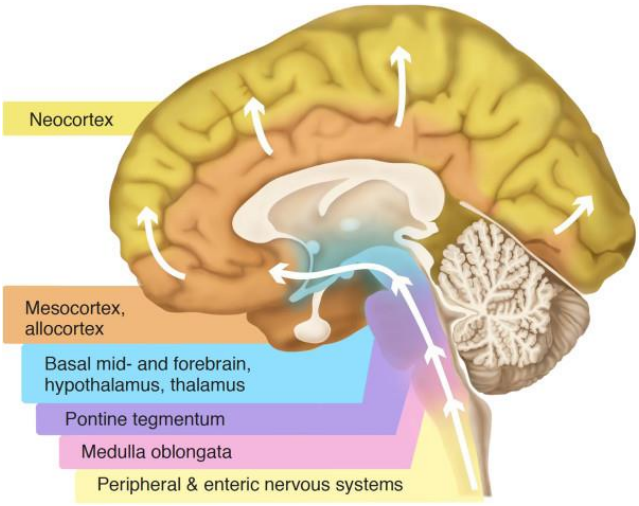

Figure 2.2: Progression of Parkinson's disease as proposed by Braak et al. From Visanji et al. 2013.

There is no cure for Parkinson's disease. Current therapies symptoms are mainly symptomatic and include pharmacologic therapies that target the motor features of PD, by improving dopamine signalling. Anticholinergics are also effective for patients with a tremor-predominant phenotype, as well as deep brain stimulation with microelectrodes, which can alleviate motor fluctuation (Erkkinen, Kim and Geschwind 2017).

2.3. Overlap between ageing and neurodegeneration

Age is the primary risk factor for AD and PD (Yankner, Lu and Loerch 2008). The overlap between ageing and neurodegeneration has fuelled a continuous debate as to whether we are all liable to develop a neurodegenerative disorder if we live long enough. This hypothesis is sustained by the fact that the aged brain presents several lesions that are not present in the brains of younger people, and that these lesions resemble a lower grade version of the ones found in neurodegenerative disease. In fact, many of the mechanisms implicated in neurodegenerative disease are parallel to the changes that occur with ageing, and the majority of aged brains show characteristic alterations, such as plaque accumulation, that can be linked to a certain level of neurodegeneration (Yankner, Lu and Loerch 2008; Hindle 2010). Moreover, the machinery behind protein synthesis, folding, disaggregation and degradation can be compromised with ageing, having profound consequences for disease presentation and progression (Wyss-Coray 2016).

Nevertheless, a minority of the population surpasses the age of 80 without showing any signs of cognitive impairment, and are designated as the “Super Agers” (Gefen et al. 2019). Super Agers highlight the possible existence of compensatory mechanisms, which are lost in normal agers, and that enable them to age healthily and maintain normal cognition.

2.3.1. Neuro-inflammageing

Inflammageing is a chronic low-grade inflammation state associated with ageing. In the CNS, this inflammatory phenotype is named neuro-inflammageing and is present in most neurodegenerative diseases, including AD and PD (Wyss-Coray 2016). Neuro-inflammageing consists of a dysregulation of the immune system, related to the activation and hyperreactivity of microglia, and the subsequent production of pro-inflammatory cytokines (Wyss-Coray 2016). There's also an increase in the number and density of microglia in the aged brain (Figure 2.3), which may be due to compensatory mechanisms, related with their reduced capacity of phagocytosis, after being engorged with debris of degenerating myelin (Gefen et al. 2019).

With ageing, astrocytes also become reactive, showing increased expression of glial fibrillary acidic protein (GFAP) (Palmer and Ousman 2018; Hol and Pekny 2015). Pro-inflammatory activated microglia and reactive astrocytes are also present in AD and PD pathogenesis. This phenomenon occurs together with neuronal loss and increase in microglia proliferation (Figure 2.3A-2.3F) (Wyss-Coray 2016; Qian and Flood 2008; Joe et al. 2018; Clayton, Van Enoo and Ikezu 2017).

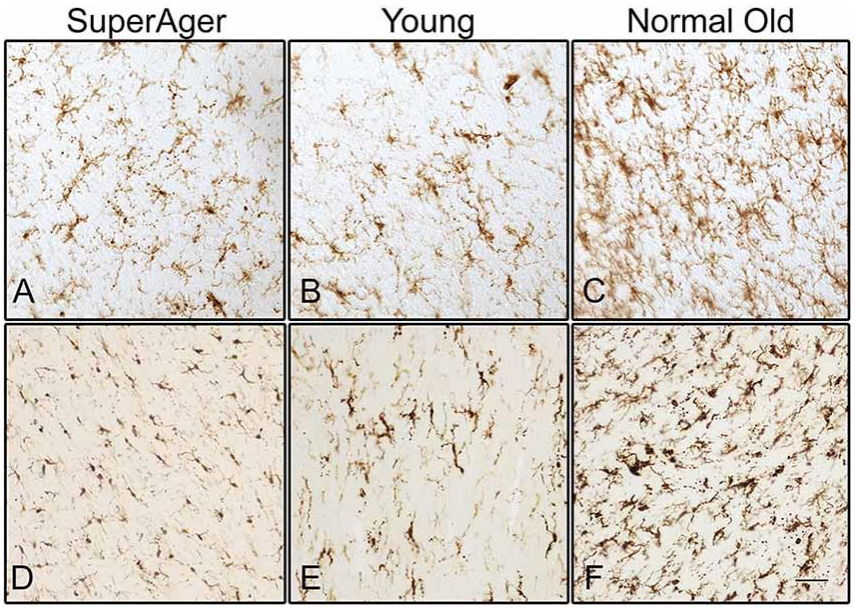

Figure 2.3: Staining of activated microglia in the white matter of the inferior frontal gyrus (A, B, C) and in corpus callosum (D, E, F). SuperAger – individual over the age of 85 whose memory test scores were at a level equal to or superior than scores of 50-to-65-year-olds. Normal Old - cognitively normal elderly over the age of 70; Young – cognitively normal young adult under the age of 24. From Gefen et al. 2019.

Activated microglia may also be involved in the generation of neuritic plaques seen in AD, either by secretion of Aβ or through the release of agents such as iron, which aggregates soluble Aβ fibrils (Wyss-Coray 2016). The progressive accumulation of iron is characteristic of ageing, and is exacerbated in selective brain regions in neurodegenerative disease, including the *substantia nigra*, leading to neurotoxicity (Ward et al. 2014; Hindle 2010).

Altered astrocytes contribute to a dysfunctional BBB with increased permeability (Yamazaki and Kanekiyo 2017), allowing activated peripheral blood immune cells, such as monocytes, to reach the

CNS. Monocytes differentiate into pro-inflammatory M1 macrophages, which produce cytokines and ROS, contributing to neuronal damage and death (Constantini, D'Angelo and Reale 2018).

In summary, neuro-inflammageing is a potential trigger of the functional changes that occur during normal and pathological ageing, and has a powerful effect on enhanced susceptibility to neurodegenerative diseases. Enhanced neuroinflammatory processes lead to loss of homeostasis in the brain environment, and, consequently, result in the death or dysfunction of neurons as seen in neurodegenerative disease.

3. Breast Cancer

Ageing is a major risk factor for the development of solid cancers (Zinger, Cho and Ben-Yehuda 2017; WHO 2011), with over 75% of all invasive cancers occurring in susceptible populations aged 55 years or older (Benz 2008). Moreover, the number of new cancer cases per year is expected to rise to 23.6 million by 2030 (National Cancer Institute 2019).

Breast cancer is also a disease primarily of older women, rarely occurring before age 30 and with the highest rates over 60 years of age (Fundo IMM Laço 2019). Over 1.5 million women (25% of all women with cancer) are diagnosed with breast cancer every year, making it the most common cancer in women worldwide and the second leading cause of cancer mortality in women (Sun et al. 2017).

The majority of breast cancers are of epithelial origin (carcinoma) and start from ductal hyperproliferation, with malignant connective tissue tumours (sarcomas) forming a negligible minority. Breast carcinomas have the potential to metastasize and can commonly transfer to distant organs such as the bones, liver, lung and brain, which mainly accounts for its mortality (Fentiman and D'Arrigo 2004; Sun et al. 2017).

Traditional breast cancer therapies include chemo- and radio-therapy. However, the treatment of breast cancer still remains a challenge due to its heterogeneous nature, which needs to be accounted for when choosing therapeutic options. Although more effective and individualized approaches to breast cancer treatment have been developed, in the past few years, there are still no targeted drugs approved for the most aggressive subtype — triple negative breast cancer (TNBC). The emergence of drug resistance also poses a threat to the successful therapy in molecular subtypes of breast cancer (BC) (Tong et al. 2018). It is therefore imperative to explore and design novel therapy alternatives, such as immunotherapy, considering that the role of the immune system in the emergence of breast cancer has been firmly established (Makhoul et al. 2018).

3.1. Breast cancer subtypes

Breast cancer is a heterogenous disease and can be classified into different subtypes with distinct biological features and different clinical implications (Dai et al. 2015). The subtyping of breast cancer has been performed based on histopathology, immunohistochemistry (IHC) markers and gene expression profiles, through the use of microarrays.

3.1.1. Histopathology

The histopathologic subtyping of breast carcinoma is based on its cytoarchitectural characteristics. Considering that it occurs in the mammary gland, the carcinoma is classified as *in situ* if it has not breached the epithelial component of the breast. Otherwise, if the carcinoma has invaded the breast stroma, it is considered invasive. Moreover, the carcinoma can be classified as ductal, if it originates from the inner lining epithelium of the ducts and lobules of the mammary gland, or lobular, if it arises from within the lobules that supply the ducts with milk (Dai et al. 2015). The combination of these features originates four different subtypes: invasive ductal carcinoma, ductal carcinoma *in situ*, invasive lobular carcinoma, and lobular carcinoma *in situ*. Invasive ductal carcinoma is the most common form of breast cancer, accounting for 50% to 70% of invasive breast cancers, while invasive lobular carcinoma accounts for 10% and is more likely to escape detection on mammography and physical examination (Alkabban and Fergunson 2019).

3.1.2. Immunohistochemistry markers

Breast cancer can be classified according to classical IHC markers, namely the oestrogen receptor (ER), progesterone receptor (PR) and the Human Epidermal growth factor Receptor 2 (HER2). Approximately 70% of all breast cancers are ER positive, meaning that oestrogen is able to bind to the ER in tumour cells, stimulating their division. Hence, ER positive tumours tend to respond well to endocrine therapy, using ER antagonists and oestrogen-producing enzymes inhibitors (Lumachi et al. 2013; Lange and Yee 2008). PR is expressed in over two-thirds of ER positive breast cancers and also has a role in the proliferation of tumour cells, in response to progesterone (Lim, Palmieri and Tilley 2016; Lange and Yee 2008). HER2 is an oncogene expressed in 20% to 30% of breast cancers, being associated with more aggressive cancers with higher recurrence and death rates. This receptor is established as a therapeutic target in HER2 positive patients, frequently using Trastuzumab, a monoclonal antibody, as a chemotherapy adjuvant (Mitri, Constantine and O'Regan 2012).

TNBC subtype does not express any of the aforementioned markers and therefore its treatment does not benefit from endocrine therapy. TNBC is typically an aggressive cancer, associated with poor prognosis and accounting for approximately 10% to 20% of all invasive breast cancers (Aysola et al. 2013). This cancer subtype does not respond well to current cancer immunotherapies and is treated almost exclusively with surgery, chemo- and radio-therapy (Feher 2017).

3.1.3. Gene expression profiling

Gene expression patterns of breast carcinomas were first assessed in a pioneering study by the Sørlie group (Sørlie et al., 2001). Using microarrays, they were able to characterize breast carcinomas at the molecular level, identifying five subtypes with distinct clinical outcomes: luminal A, luminal B, HER2 overexpression, basal and normal-like tumours. These subtypes are also associated with the IHC nomenclature, with exception of the latter.

Luminal A breast cancer is the most common subtype, accounting for approximately 50% of all invasive breast cancers (Makki 2015). It is usually also positive for ER, PR and HER2. Luminal B breast cancer is the second most common subtype, accounting for 20% of all invasive breast cancers (Makki 2015).

Like luminal A, it is positive for ER and PR. However, only part of these tumours express HER2, and are usually associated with worse prognosis, relative to luminal A. In general, luminal tumours are associated with good prognosis, responding well to endocrine therapy but poorly to conventional chemotherapy (Dai et al. 2015).

HER2 overexpressing tumours are negative for both ER and PR, and positive for HER2. These tumours, which are identified at the transcriptomic level, do not perfectly match tumours positive for HER2, identified through IHC. This subtype is associated with poor prognosis, mainly due to the fact that patients have a higher risk of relapsing (Dai et al. 2015).

Basal-like is a breast cancer subtype that has a tendency to affect younger women. It is typically a TNBC, and, although the latter comprises a more heterogenous group of tumours, there are studies defending that basal-like tumours should be divided into further categories, considering that patients may have divergent outcomes regarding mortality rates or recurrence (Milioli et al. 2017; Liu, Zhang and Zhang 2014). Like TNBC, the lack of hormone receptors leads to a limitation in therapeutic options, showing a lack of response to Trastuzumab and endocrine therapies. Thus, this subtype is usually associated with poor prognosis (Liu, Zhang and Zhang 1014; Makki 1015).

Normal-like tumours are TNBC distinct from basal-like, with an intermediate prognosis between the latter and luminal cancers. This subtype is poorly described and its clinical relevance is yet to be determined, due to the fact that it could be the result of technical artefacts from normal breast tissue contamination (Yersal and Barutca 2014; Wesolowski and Ramaswamy 2018).

3.2. The tumour microenvironment

The tumour microenvironment (TME) is an important determinant of the initiation, progression and metastasis of cancer. It consists of a complex and dynamic niche, made of extracellular matrix and cellular components. The main infiltrating cells include fibroblasts, neuroendocrine, adipose, and immune cells. Immune cells in the TME are both from the innate and immune system (Wang et al. 2017).

Despite the existence of different breast cancer subtypes and patient inter-variability, recent studies focused on the TME have been unveiling tumour-infiltrating immune cell patterns that are either correlated with poor patient prognosis or good outcome (Man et al. 2013; Fridman et al. 2010). One example of this relationship is the association between the infiltration of lymphocytes in breast tumours and prognosis. Amongst tumour-infiltrating lymphocytes (TILs), NK cells are able to naturally kill circulating tumour cells through cytolytic activity (Wu and Lanier 2003). However, while the role of NK cells outside the vascular system and in the tumour microenvironment still remains unclear in most cancer types (Larsen, Gao and Basse 2014), the association between high levels of CD8+ T cells and their anti-tumour effect is well established (Maimela, Liu and Zhang 2019). On the other hand, regulatory T (Treg) cells present immunosuppressive functions that are able to inhibit the cytotoxic activity of CD8+ T cells and NK cells, thus promoting tumour growth (Maimela, Liu and Zhang 2019).

Tumour-infiltrating T cells have already been described as a key component of the tumour microenvironment, due to recent clinical trials that demonstrated the ability to durably control cancer in some patients by manipulating them (Sharma and Allison 2015). Tumour-associated myeloid cells, however, remain less studied than T cells, even though they comprise a significant part of the total

tumour immune cell infiltrates and have already been described as potent regulators of tumour-associated immune suppression, cell invasion and metastases (Weagel et al., 2015).

During tumour development, tumour cells secrete several cytokines, e.g., monocyte colony stimulating factor (M-CSF) and granulocyte-macrophage colony-stimulating factor (GM-CSF) (Tang et al. 1992; Kerkar and Restifo 2012). GM-CSF is a key factor in the differentiation of myeloid-derived suppressor cells (MDSCs). MDSCs are myeloid cells of granulocytic (neutrophil-like) or monocytic (monocyte-like) origin, induced into an immature and suppressor state. These cells are capable of not only supressing NK and CD8+ T cell anti-tumour activity, but also directly stimulating tumour cell proliferation, metastasis and angiogenesis (Kerkar and Restifo 2012; Mabuchi et al. 2018; Gonda et al. 2017). Moreover, GM-CSF, together with M-CSF, induce the recruitment of monocytes and their differentiation into non-polarized (M0) macrophages (Ushach and Zlotnik 2016; Qiu et al. 2018; Tang et al. 1992).

Tumour associated macrophages (TAMs) are a major component of the tumour microenvironment, reaching over 50% of the tumour mass in some breast cancers (Weagel et al., 2015; Qiu et al. 2018). Depending on the stimuli from the surrounding TME, macrophages are polarized into classically activated (M1) or alternatively activated macrophages (M2) and exert dual influences on tumourigenesis by either enhancing anti-tumour responses or by manifesting tumour-promoting activities, respectively. M2 macrophages have already been described as being potent regulators of tumour-associated immune suppression, cell invasion and metastasis. M2 macrophages aid in the process of angiogenesis, allowing new blood vessel growth, which feeds the malignant mass of cells (Weagel et al., 2015). Thus, their presence in tumour masses can be an indicator of poor prognosis in numerous cancer types, including breast cancer (Weagel et al., 2015; Cotechini, Medler and Coussens 2015; Gonda et al. 2017).

On the other hand, M1 macrophages are able to induce tumour-regression, through contact-dependent phagocytosis and cytokine production, and are usually associated with good prognosis (Poh and Ernst 2018). The contrasting roles of TAMs in breast cancer makes them an active topic of research, with the prospective of using these cells as targets in cancer immunotherapy, either by reducing the numbers of M2 macrophages and/or inducing re-polarization towards a M1 phenotype (Poh and Ernst 2018).

3.2.1. Immunosenescence and macroph-ageing

Throughout the course of life, the immune system keeps surveilling the organism for foreign pathogens, such as bacteria and viruses, and pre-cancerous and cancerous cells. This is a process known as immunosurveillance (Smyth, Dunn and Schreiber 2006). However, there is a gradual decline of the immune system with age, which is defined as immunosenescence (Zinger, Cho and Ben-Yehuda 2017). This deterioration leads to alterations in the proportion of different immune cell types in the organism and in their capabilities, including a decrease in the number of naïve CD8+ and CD4+ T cells (Pawelec 2017), impaired function of mature lymphocytes, and a decrease in the number and the phenotypic alteration of natural killer (NK) cells (Zinger, Cho and Ben-Yehuda 2017; Montecino-Rodriguez, Berent-Maoz and Dorshkind 2013). There is also evidence for an increase in myeloid-lineage cells (Montecino-Rodriguez, Berent-Maoz and Dorshkind 2013), which is coupled with reduced anti-tumour activity in macrophages and their predisposition for a pro-tumour phenotype (Figure 3.1) (Provinciali et al. 2017).

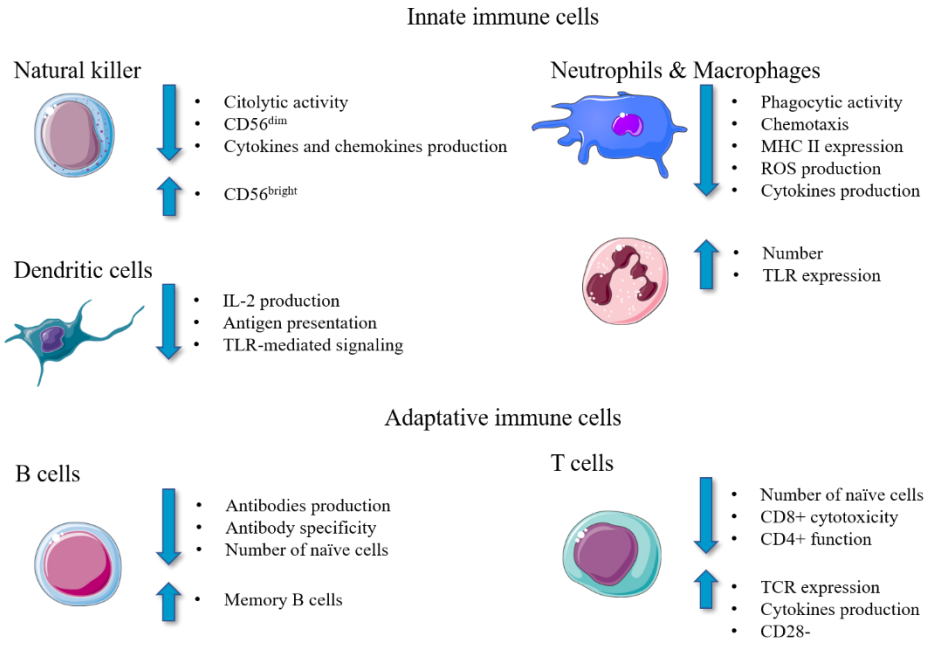

Figure 3.1 Schematic representation of immune cell immunosenescence-related changes. Adapted from Costantini, D'Angelo, and Reale 2018.

Macroph-ageing was defined by the Franceschi group (Franceschi et al. 2006) as the chronic macrophage activation, one of the major factors responsible for the pro-inflammatory state associated with ageing. Macroph-ageing and inflamm-ageing occur in association with immunosenescence, and lead to the reduction in the efficacy of the immune response, resulting in decreased immunosurveillance and anti-tumour effector function (Costantini, D'Angelo, and Reale 2018).

Increasing breast cancer rates with ageing have already been associated with immunosenescence. This association might not only impact the incidence and progression of breast cancer, but also the effectiveness of immunotherapy due to compromised immunosurveillance (Provinciali et al. 2017). The efficacy of more traditional cancer therapies, such as chemo- and radio-therapy, may also be impacted, given that they are at least partially dependent of tumour-specific immune responses and their ability to bring back immunosurveillance (Bracci, Schiavoni and Sistigu 2013; Kroemer et al. 2015). Therefore, the characterization of tumour-infiltrating immune cells may disclose better strategies for overcoming immune suppression and restoring immunosurveillance.

4. Single-cell transcriptomics

In the 17th century, Robert Hooke first used a microscope to describe little boxes distinct from one another in cork, which he named cells (Hooke 1665). Today, it is universally accepted that the cell is the unit of life, all cells come from pre-existing cells and all living organisms are composed of different cell types that share common basic features, but may vary vastly in their function and molecular profile (Feher 2017).

It is important to assess this cell heterogeneity in order to understand fundamental biological processes in health and disease, allowing for the improvement of existing therapies and the discovery of new ones (Hwang, Lee and Bang 2018). One way to do this is through single-cell RNA sequencing (scRNA-seq).

scRNA-seq was first described in 2009 by Tang et al. and has since been gaining widespread popularity as a method to survey the diversity of cell types within a tissue sample. Its underlying principles are the same as in bulk RNA-sequencing (RNA-seq), an approach that uses sequencing technologies to profile the transcriptome, i.e., the complete set of RNA molecules in a biological sample, thereby allowing to quantify and assess differential gene expression between conditions of interest (Wang, Gerstein and Snyder, 2009). However, common bulk RNA-seq experiments measure gene expression levels as averages across populations of cells, under the assumption that tissues are composed of homogeneous populations of cells, only allowing the characterization of population-level gene expression. These methods are therefore insufficient for studying heterogeneous cellular systems due to the likelihood of missing important cell-to-cell variability (Hwang, Lee and Bang 2018). scRNA-seq experiments measure the distribution of expression levels for each gene in each cell, allowing to characterize a population of individual cells and to better understand gene expression patterns in complex heterogeneous tissues (Wang et al. 2018).

Although the development of scRNA-seq protocols was motivated by the need to study conditions where only a small amount of material was available, such as cells from the early embryonic development, protocol improvements and massively parallel sequencing platforms boosted an increase in the number of cells studied in these analyses, ranging from only a few cells to hundreds of thousands of single cells per study (Figure 4.1) (Tang et al. 2009; Svensson, Vento-Tormo and Teichmann 2018). This continuous evolution is radically improving the dissection of heterogeneity within cell populations, with many applications in diverse fields. One example is the deconvolution of highly diverse immune cell populations in health and disease (Hwang, Lee and Bang 2018; Papalexi and Satija 2017), which has been applied in the cancer research field, by enabling the estimation of specific immune cell composition of solid tumours (Schelker et al. 2017), and in the neurosciences research field, by allowing the characterization of neuronal and non-neuronal cell diversity across multiple human brain regions (Spaethling et al. 2017; Darmanis et al., 2015).

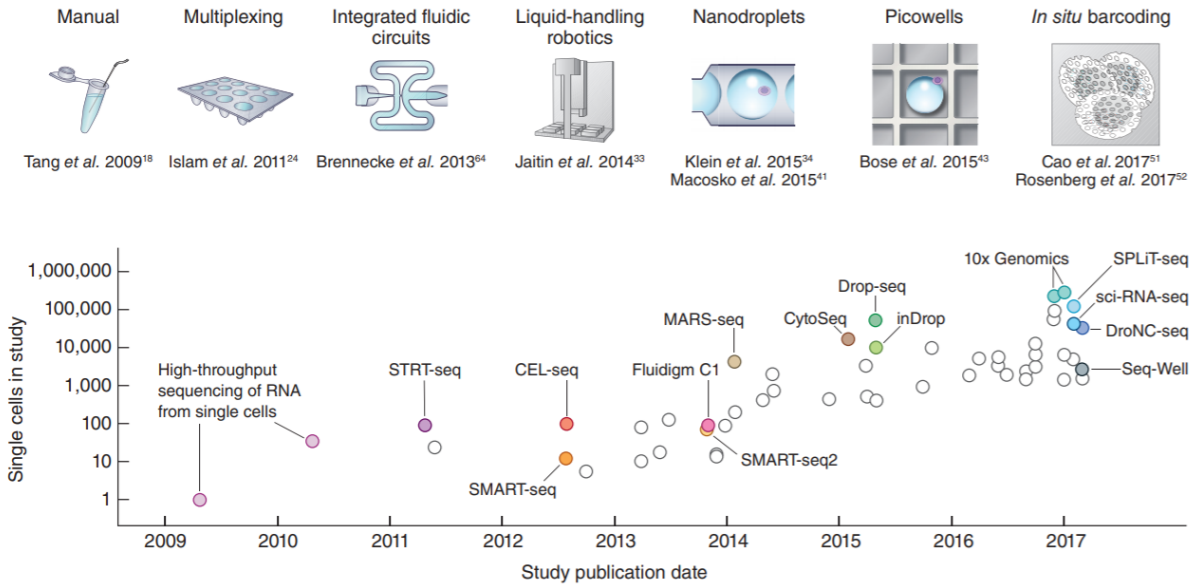

Figure 4.1: Scaling up of scRNA-seq experiments. From Svensson, Vento-Tormo and Teichmann 2018.

4.1. The technology of single-cell RNA sequencing

Every scRNA-seq experiment shares a common basic strategy, starting with tissue dissociation, sorting and isolation of single cells, followed by reverse transcription of mRNA, amplification of cDNA and, finally, library preparation and sequencing. Like in bulk RNA-sequencing, most protocols use the Illumina platform for sequencing. Thus, the major difference between bulk RNA-seq and scRNA-seq protocols is the sorting step.

There are several methods that can be used to sort cells, including the ones that capture only a small number of cells with high level of supervision, such as micromanipulation and laser capture microdissection (LCM), and automatic microfluidic methods which focus mainly on high throughput (Figure 4.2).

Micromanipulation consists of using microscope-guided capillary pipettes to extract single cells, while LCM uses a laser to attach individual cells from a tissue to a thin film that can be eliminated afterwards. These low throughput methods have the advantage of ensuring that only one cell is captured at a time, and allow the verification of cell viability and morphologic features through microscopic supervision (Kolodziejczyk et al. 2015).

Microfluidic platforms, on the other hand, allow sorting hundreds of single cells in each run. They consist of using integrated fluidic circuits to capture cells and isolate them in individual channels. Cell capture is followed by automated reverse transcription and pre-amplification in nanoliter volumes, which allows to control for reagent costs. One of the most common microfluidic platforms is the Fluidigm C1 (Gong, Do and Ramakrishnan 2018). However, this method possesses several restrictions, such as the limited size of the capturing sites, which implies that cells must also be relatively homogenous in size, and the low yield of captured cells, considering the need to input at least 1,000 cells to recover 96 per chip. To surpass these limitations, micro-droplet-based microfluidic platforms were developed. These do not use capturing sites, but instead encapsulate cells in oil droplets, together with the reagents needed for reverse transcription and amplification. Micro-droplet technologies, such as the 10X Genomics chromium system (See et al. 2018), are gaining popularity in the scRNA-seq field, due to escalating captured cell numbers up to thousands of cells per run, while also increasing yield up to 50% (Kolodziejczyk et al. 2015; Gong, Do and Ramakrishnan 2018).

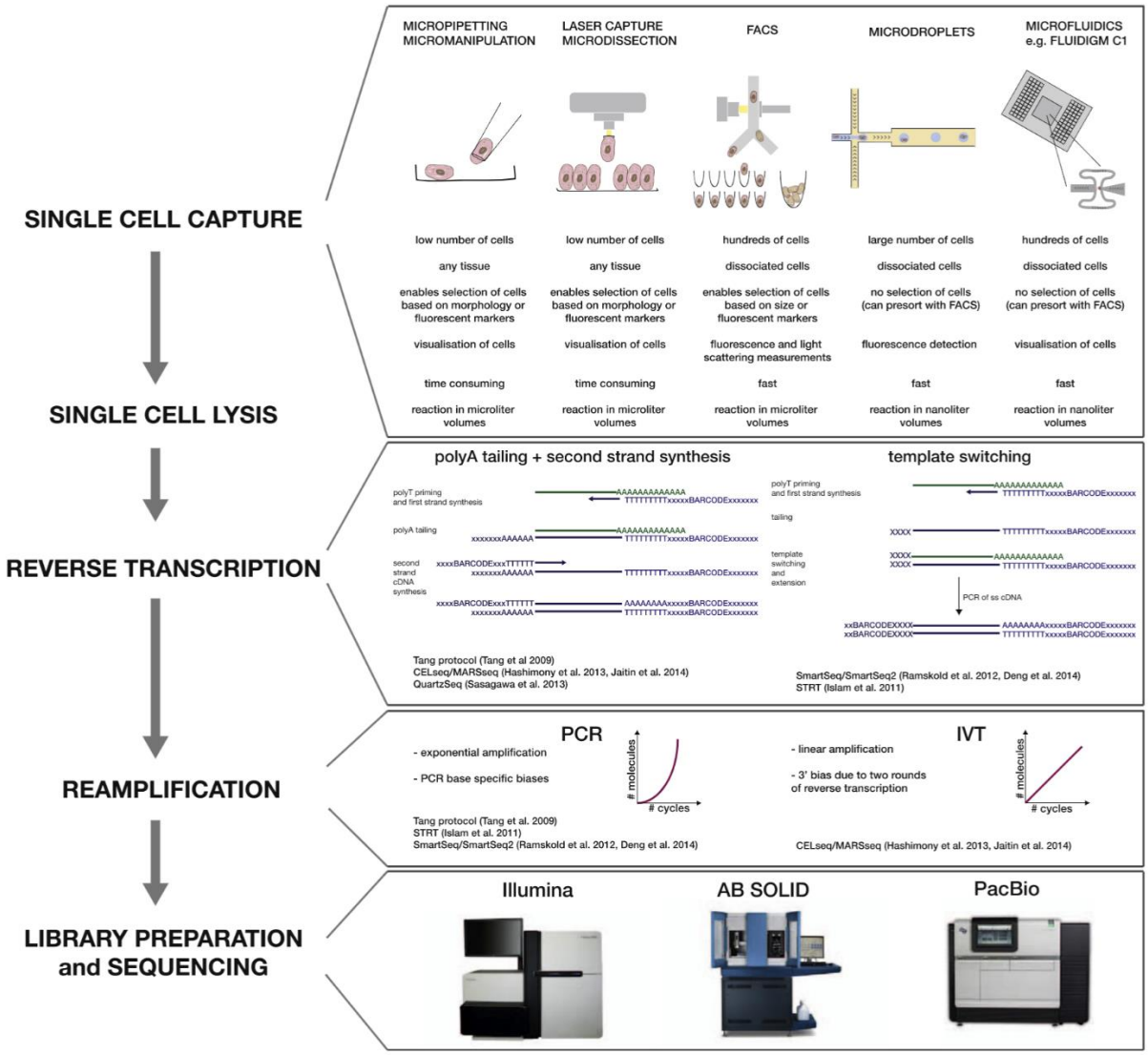

Figure 4.2: Steps of a scRNA-seq protocol with different experimental approaches. From Kolodziejczyk et al. 2015.

4.1.1. Number of cells vs sequencing depth

The high number of samples used in scRNA-seq means that these assays usually involve a trade-off between the number of cells and sequencing depth, i.e., the number of RNA transcripts sequenced per cell. There is a direct dependence of cell type classification on sequencing depth, meaning that the resolution with which we are able to distinguish different cell types is dependent on the depth of cellular profiling (Streets and Huang 2014).

Previous systematic analyses of how the transcriptional identity of a cell is preserved, as sequencing depth is decreased, concluded that a majority of the primary genes that contribute to transcriptional variance among diverse types of cells are identified by low-coverage sequencing analysis, preserved in sequencing depths as low as 10,000 reads per cell. Cells with subtle transcriptional differences, such as neural cells at different stages of development, are distinguishable with a sequencing depth of about 50,000 reads per cell (Figure 4.3) (Streets and Huang 2014).

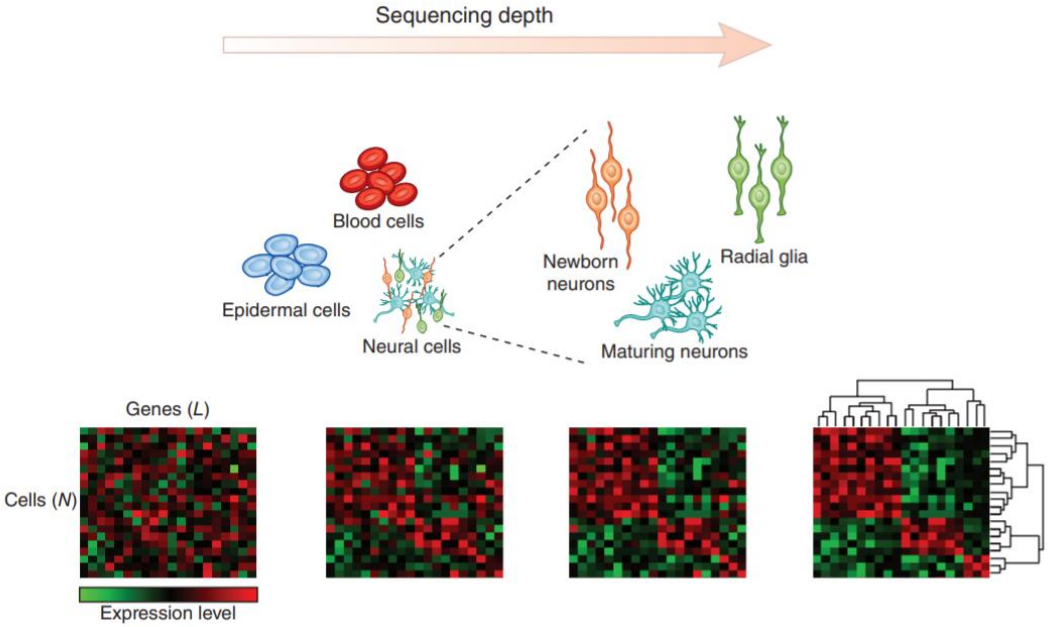

Figure 4.3: Relationship between sequencing depth and cell type identification. Resolution of different cell types and states increases with sequencing depth. Cells from different organs can be distinguished with less than 10,000 reads, while cells from the same type and at different stages of development can be identified with about 50,000 reads per cell. Adapted from Streets and Huang 2014.

Nevertheless, accurate estimation of sequencing depth per experiment should take into account both total mRNA content in individual cells and the diversity of mRNA species in those cells. Different scRNA-seq protocols with different cell number and sequencing depth capacities can be used in an integrated approach, whereby complex tissues are first discriminated using low-depth droplet-based technologies to identify new and/or rare populations of interest and their associated markers. Subsequently, these markers can be used for enrichment and deep sequencing using high-depth and low throughput approaches (Baran-Gale, Chandra and Kirschner 2017; Papalexi and Satija 2018).

4.2. Computational analysis

In scRNA-seq, each library represents a single cell rather than a population of cells, as is the case for bulk RNA-seq. Working with minute amounts of mRNA means that only 10%–20% of transcripts in a cell get reverse transcribed, leading to high technical noise, especially noticeable in lowly expressed genes (Kolodziejczyk et al. 2015).

The sources of discrepancy between libraries arise mostly from amplification bias and gene dropouts. Dropouts occur when a gene is expressed at moderate level in one cell, but fails to be captured in another cell, usually due to failures in reverse transcription. This binary modality in gene expression associated with zero-inflation is usually modelled in methods developed specifically for scRNA-seq data analysis (Sekula, Gaskins and Datta 2019).

Given that scRNA-seq is a relatively recent technology, currently there is still no gold standard pipeline to analyse this type of data. This leads to an underlying challenge in scRNA-seq data analysis, considering the vast variability between different datasets regarding experimental design, number of

cells and sequencing depth. Therefore, scRNA-seq pipelines should be adapted for each individual dataset in terms of quality controls measures and statistical tools.

However, it is possible to pre-define a few typical sequential steps in scRNA-seq data analysis. These include quality control to remove poor quality samples, followed by normalization for library sizes, correction of technical confounders, and clustering to identify groups of cell types and/or states (Kolodziejczyk et al. 2015).

Dimensionality reduction techniques are also commonly used to perform exploratory analysis of scRNA-seq data, as these data are typically high-dimensional. Two of the most common dimensionality reduction techniques are principal component analysis (PCA) (Jolliffe and Cadima, 2016) and t-distributed stochastic neighbour embedding (t-SNE) (van der Maaten and Hinton, 2008).

In PCA, the original variables are substituted by a new set of uncorrelated variables. These are designated as principal components, and consist of linear combinations of the original variables. Of all the possible linear combinations, for each case, the one with the maximum variance is selected, given that the principal components should explain a major part of the variance associated with the original variables. Hence, the variance of each principal component is a measure of the quantity of information explained by it. Dimensionality reduction is obtained by considering only the principal components that are associated with the higher variance, resulting in small information loss (Jolliffe and Cadima, 2016).

t-SNE, on the other hand, is a nonlinear dimensionality reduction. This method consists of defining a similarity probability distribution over pairs of objects in high-dimensional space, and calculating the probability of similarity of points in the corresponding low-dimensional space. It then tries to minimize the difference between similarities in high- and low-dimensional space, by using the Kullback-Leibler divergence (Pavlichin and Weissman, 2016), a measure of how one probability distribution diverges from a second expected probability distribution. This results in the mapping of the data to a lower dimensional space and allows the identification of clusters based on similarity of data points with multiple features. However, unlike PCA, the input variables are no longer identifiable (van der Maaten and Hinton, 2008).

The remaining downstream analyses vary according to the biological question under study. For example, one might be interested in finding marker genes that distinguish different cell types from each other, for which a classic differential expression analysis suffices (Alessandrì, Arigoni and Calogero 2019). Another possibility would be to study the alteration of states of a given cell type between two conditions, e.g., microglia in healthy and microglia in AD-affected brains, which would refer us to differential state analysis (Crowell et al. 2019). In this work, we focused on constructing scRNA-seq data analysis pipelines by combining the best approaches and necessary quality metrics for each analysed dataset, with the general goal of finding marker genes of different cell types and performing cell type deconvolution of bulk RNA-seq datasets.

5. Objectives

5.1. Single-cell RNA sequencing of the brain

The existing knowledge on the extent of cell type diversity in the mammalian brain remains incomplete (Tasic et al. 2016), as remains the estimation of the relative cell type abundance in the nervous system, which is considered an important approach to understand neurological disease and ageing. Indeed, many studies have described a large number of neurological diseases implicating abnormal glial cell numbers (von Bartheld, Bahney, and Herculano-Houzel 2016) and also a loss or functional alteration of astrocytes and microglia associated with ageing, which is itself an important risk factor for the onset and progression of neurodegenerative disorders, such as Parkinson's and Alzheimer's diseases (Joe et al. 2018; Keren-Shaul et al. 2017; González-Reyes et al. 2017).

As shown by this project's host laboratory at iMM, led by Nuno Barbosa Morais, through a preliminary gene expression analysis of *post-mortem* prefrontal cortex samples of sporadic PD and non-PD individuals, there is an expected loss of neurons and a specific accentuation in glial cells of the systemic molecular effects of the disorder reported for brain tissues (activation of inflammation and immune response pathways and lower activity of metabolic ones).

Therefore, our first goals are to:

a. Obtain the gene expression signatures of the major brain cell types;
b. Unveil how the relative abundance of neurons and glia in the brain correlate with ageing and neurological health.

5.2. Single-cell RNA sequencing of tumour-infiltrating immune cells

There is an urge to characterise the anti-cancer myeloid compartment of the tumour microenvironment and understand the complexity of the mechanisms involved in myeloid regulation (Elliot et al. 2017). At the same time, age-associated changes in the tumour-infiltrating immune landscape and the consequence of immunosenescence are still poorly described in breast cancer. Moreover, most immunotherapy pre-clinical experiments are performed in young mice, leading to the suggestion that successful treatments may be age-dependent and biased towards younger subjects (Zinger, Cho, and Ben-Yehuda 2017).

All of these considered, the second overall aim of this project is to unbiasedly characterise cellular heterogeneity in the tumour microenvironment, focusing on the context of breast cancer, while also unveiling the variability of infiltration patterns with age and prognosis.

Particularly, we aim at:

a. Evaluating the cellular diversity and molecular signature of TAMs;
b. Understanding how the intra-tumoural diversity and functionality of infiltrating immune cell types is associated with age and prognosis.

6. Methods

R programming language (ver. 3.6.1) (R Core Team 2019), with several Bioconductor packages (ver. 3.9) (Huber et al., 2015) and packages from The Comprehensive R Archive Network (CRAN), was used to perform most downstream steps, including quality control, visualization, normalization, feature selection, clustering and differential expression analysis.

6.1. Single-cell RNA sequencing of the brain

6.1.1. Datasets

Table 6.1: Summarized description of the datasets used in the first part of the analysis.

ID	Manuscript designation	Database	Technology	Format	Samples
phs000835.v7	Spaethling	dbGAP*	scRNA-seq	FASTQ	506
GSE67835	Darmanis	GEO**	scRNA-seq	Count matrix	247
GSE73721	Zhang	GEO	scRNA-seq	Count matrix	29
phs000424.v7.p2	GTEx	GTEx***	RNA-seq	Count matrix	2,467
GSE125583	Alzheimer's disease	GEO	RNA-seq	Count matrix	289
GSE68719	Parkinson's disease	GEO	RNA-seq	Normalized count matrix	57

* database of Genotypes and Phenotypes (Mailman et al. 2007).

** Gene Expression Omnibus (Clough and Barrett 2016).

*** The Genotype-Tissue Expression (Lonsdale et al. 2013).

The phs000835.v7 dataset, here referred as the **Spaethling dataset**, consists of healthy tissue samples from adult human brain biopsies collected from patients at the time of surgery and then further cultured up to 84 days *in vitro*. Biopsies were performed in different brain areas: hippocampus, left and right frontal cortex, left temporal lobe, right middle temporal gyrus, left primary motor cortex and right cerebellum. Single-cells were sorted using microcapillary pipette aspiration and snap frozen until processing. Each single-cell's RNA was amplified and prepared for sequencing, following the TruSeq stranded library generation without the initial fragmentation step. Samples were deep paired-end sequenced with either an Illumina HiSeq 2500 or a NextSeq 500 machine. Out of the 506 cells, 73 were labelled by the authors as astrocytes, 30 as microglia, 136 as neurons, 4 as oligodendrocytes, 38 as endothelial cells and 229 samples were left as unidentified brain cells. We extracted the paired FASTQ-files from the SRA-raw files using the *fastq-dump* program of the SRAtoolkit (Leinonen, Sugawara and Shumway, 2010). We then quantified transcript abundances against a reference transcriptome using the pseudoaligment tool Kallisto (Bray et al. 2016).

The GSE67835 dataset, here referred as the **Darmanis dataset**, consists of healthy tissue samples from 8 adult human brain cortex biopsies. Single cells were sorted using the Fluidigm C1 platform. Each single-cell's RNA was amplified and prepared for sequencing using the Nextera XT DNA Sample Preparation Kit. Samples were sequenced with an Illumina NextSeq instrument (2x75 bp) with an

average of 2,838,000 reads per cell. Reads were aligned using STAR (Dobin et al. 2012) and per gene counts were calculated using HTSEQ (Darmanis et al. 2015). Quality control of the Darmanis dataset was previously performed in-house, using the same quality control metrics applied in this work. Out of 247 cells, 62 were labelled as astrocytes, 16 as microglia, 131 as neurons and 38 as oligodendrocytes.

The GSE73721 dataset, here referred as the **Zhang dataset**, consists of healthy tissue samples from adult human and mouse brain temporal lobe cortex biopsies. Single cells were sorted using micromanipulation coupled with cell type specific antibodies (Zhang et al. 2016). Each single-cell's RNA was amplified and prepared for sequencing using the Next Ultra RNA-seq library prep kit for Illumina. Libraries were sequenced using the Illumina NextSeq sequencer (2x150 bp). Out of the 29 cells used in this project, 12 were labelled by the authors as human mature astrocytes, 3 as myeloid cells (microglia), 4 as mouse mature astrocytes, 1 as human mature neuron, and 5 as human mature oligodendrocytes (Zhang et al., 2016).

The phs000424.v7.p2 dataset, here referred as the **GTEx dataset**, was obtained from the GTEx project, a resource database and associated tissue bank with relevant clinic metadata (The GTEx Consortium, 2013). This dataset consists of 2,467 bulk RNA-seq samples from several neural tissues, namely, amygdala, anterior cingulate cortex, caudate (basal ganglia), cerebellar hemisphere, cerebellum, cortex, frontal cortex, hippocampus, hypothalamus, *nucleus accumbens* (basal ganglia), *putamen* (basal ganglia), spinal cord (cervical c-1) and *substantia nigra*. Libraries were prepared using the Illumina TruSeq RNA protocol (2x76 bp), and sequenced with the Illumina HiSeq 2000 platform, averaging ~50 million aligned reads per sample.

The GSE125583 dataset, here referred as the **Alzheimer's disease dataset**, consists of 242 samples of *post-mortem* fusiform gyrus from Alzheimer's disease patients, and 126 controls, i.e. samples of fusiform gyrus of neurologically normal age-matched controls. This dataset was available as a raw count matrix (Srinivasan et al., 2019). Reads were aligned with GSNAP (Wu et al. 2016) and reads per gene were counted with the HTSeqGenie Bioconductor package (Pau and Reeder, 2014).

The GSE68719 dataset, here referred as the **Parkinson's disease dataset**, consists of samples of *post-mortem* prefrontal cortex (BA9) tissue obtained from 29 Parkinson's disease patients and 44 healthy controls. Samples were sequenced on an Illumina HiSeq2000 instrument (2x101 bp). This dataset was available as a DESeq2 normalized count matrix (Dumitriu et al. 2015). To avoid having age as a confounding factor, we selected only samples from patients over 64 years of age, resulting in 29 PD samples and 28 non-PD samples.

6.1.2. Normalization

Normalization of library sizes of scRNA-seq datasets was performed using a deconvolution-based scaling method, available as the *computeSumFactors* function in the package scran (Lun, Bach and Marioni 2016).

Normalization for the correction of batch effects of scRNA-seq datasets was performed using the *removeBatchEffect* function, available in Bioconductor/R-package limma (Ritchie et al., 2015). This method fits a linear model to the data and removes the component related to the batch effects (Ritchie et al. 2015).

Normalization of library sizes of bulk RNA-seq datasets was performed using the voom method, which estimates the mean-variance relationship of the log-counts and generates a precision weight for each

observation. This method is available as the *voom* function in Bioconductor/R-package limma (Ritchie et al., 2015).

6.1.3. Feature selection

Feature selection, also known as variable selection, consists of selecting a subset of meaningful and relevant features, reducing data dimension before performing predictive analysis. This allows to improve efficiency in classification models, reduce computation complexity, and boost the generalization ability of the model to external datasets (Lai et al., 2017).

We tested two different methods for feature selection, identification of highly variable genes (HVG) and identification of high dropout genes (HDG).

Identification of HVG was applied using the *Brennecke_getVariableGenes* function with default parameters, included in the M3Drop package (Andrews 2019). This function corrects for the relationship between variance and mean expression by normalizing library sizes and calculating the mean and the squared coefficient of variation (SCV). HVG are selected by fitting a quadratic curve to the relationship between the mean and SCV, and using a chi-squared test to find genes significantly above the curve (Brennecke et al. 2013).

Identification of HDG was applied using the *M3DropDifferentialExpression* function with default parameters, available in the M3Drop package (Andrews 2019). This method models the non-linear relationship between dropout rates and average gene expression, using a Michaelis-Menten equation, considering that dropouts occur during reverse transcription, an enzyme reaction whose kinetics are thereby modelled. HDG are identified as being shifted above the expected curve (Andrews and Hemberg 2018).

6.1.4. Pseudotime analysis

To analyse different states of development, we used the Monocle package (Trapnell et al. 2014) to build a neural cell trajectory, in which cells were ordered by pseudotime, i.e., a measure of how much progress an individual cell has made through a process such as cell differentiation.

We began by using the function *setOrderingFilter* to mark the genes used in the analysis, in this case, the set of HDG identified using the M3Drop package. This was followed by reducing the dimensionality of the dataset, with the *reduceDimension* function, with method *DDRTree*. Finally, cells were ordered with the *orderCells* function, which calculates where each cell falls within that trajectory, using a reversed graph embedding machine learning method to learn the sequence of gene expression changes. Thus, this trajectory is associated with the total amount of transcriptional change that a cell undergoes as it moves from the starting state to the end state (Trapnell et al. 2014).

6.1.5. Clustering

We used the *sc3* function from the SC3 package (Kiselev et al. 2017) to perform unsupervised clustering of single neuronal and glia cells. The *ks* parameter (number of clusters) was defined as seven, considering this was the smallest number of *ks* for which there was separation of neurons and microglia into distinct clusters. Working as a consensus clustering tool, this method uses a parallelisation approach to evaluate different clustering parameters simultaneously, e.g., the set of Euclidean, Pearson and Spearman distances. Finally, it combines the different clustering results into a consensus matrix that demonstrates the probability of a cell belonging to a given cluster (Kiselev et al. 2017).

6.1.6. Marker genes

We tested three different methods for obtaining marker genes that could discriminate each cell subset, implemented in the PAMR (prediction analysis for microarrays) (Tibshirani et al. 2002), SC3 (Kiselev et al. 2017) and Caret (Classification And REgression Training) packages (Kuhn, 2019).

PAMR uses the nearest shrunken centroid classification, which consists of computing a standardized centroid for each class/cell type, by averaging gene expression for each gene in each class and dividing it by the within-class standard deviation for that gene. Each of the centroids is shrank towards 0, by an amount defined as the threshold. This reduces the effect of noisy genes and performs automatic selection of marker genes (Tibshirani et al. 2002). The implementation of this method follows a sequential use of the available functions, starting with the *pamr.train* function, which receives the count matrix as input and trains the nearest shrunken centroid classifier. Next, we use *pamr.cv* to cross-validate the classifier and *pamr.fdr* to estimate false discovery rates. Finally, *pamr.listgenes* outputs a list of marker genes based on the user-defined parameter *t* (threshold). We chose *t* as the value that allowed the utilization of the least number of marker genes possible, while retaining a high accuracy of classification (> 90%), along with False Discovery Rate (FDR) = 0.

The method implemented in the SC3 function *get_marker_genes* (Kiselev et al. 2017), that enables the discovery of marker genes for each cell type, consists of building a binary classifier for each gene based on the mean expression values of each cluster. The area under the receiver operating characteristic (AUROC) curve is used to quantify the accuracy of the classifier and a p-value is calculated for each gene using the Wilcoxon signed rank test. For the analysis of neuronal and glia cells, we defined an *auroc.threshold* > 0.75 and *p.val* < 0.05, with a maximum of 200 marker genes for each cell type (Kiselev et al. 2017).

The Caret package comprises a set of functions applied in predictive modelling. For this project, we focused on elastic nets, a type of regularized logistic regression (Kuhn, 2019). This method linearly combines two regularization parameters, the L1 (λ) and L2 (α) penalties of the lasso (least absolute shrinkage and selection operator) and ridge methods, respectively. High values of λ lead to many coefficients being zeroed, thus performing feature selection (Waldmann et al. 2013). We implemented this method using the *train* function, with the method parameter defined as “glmnet”.

6.1.7. Cell type deconvolution

Deconvolution algorithms make use of specific cell types' gene signatures expected to be contained in a heterogeneous tissue sample, such as is the case for the expected different cell types in brain samples, to estimate their cell fractions within the tissue. This is performed assuming that the gene expression profile of a given tissue is the convolution of the gene expression levels of the different cells that constitute it (Finotello and Trajanoski 2018).

We used CIBERSORT (Newman et al. 2019) to perform neuronal and glia cell type deconvolution of bulk RNA-seq brain samples. CIBERSORT uses linear support vector regression to characterize cell type composition of complex tissues from their gene expression profiles (Newman et al. 2019).

6.2. Single-cell RNA sequencing of tumour-infiltrating immune cells

6.2.1. Datasets

Table 6.2: Summarized description of the datasets used in the second part of the analysis.

ID	Manuscript designation	Database	Technology	Format	Samples
GSE114725	Azizi	GEO	scRNA-seq	Count matrix	21,253
-	TCGA	TCGA*	RNA-seq	Count matrix	878

* The Cancer Genome Atlas (Tomczak et al. 2015; Weinstein et al. 2013).

The GSE114725 dataset, here referred to as the **Azizi dataset**, consisted of scRNA-seq of breast tumour-infiltrating immune cells (Azizi et al. 2018). Immune cells were obtained from 8 different patients/tumours and also from patient's blood, lymph node, and normal tissue. For this project, only the tumour-infiltrating immune cells were used, comprising a total of 21253 cells. The tumours are represented across several subtypes, namely ER positive, PR positive, HER2 positive and TNBC. Single-cells were previously sorted using FACS and droplet technology, and sequenced using Illumina HiSeq 2500 instruments with paired-end sequencing (PE1 54 bp and PE2 66 bp). Each cell was covered by an average of 22,000 reads. Quality control was performed using FASTQC (Andrews, 2010). Libraries that displayed significant (> 25%) low quality bases were re-sequenced to make sure samples were comparable in the downstream analysis. The count matrix was generated using package SEquence Quality Control (SEQC) (Azizi et al. 2018). Poor quality cells were filtered using the usual quality metrics, library size, number of unique features detected and percentage of mitochondrial genes. This generated a count matrix that was the starting point of the analysis performed in this project.

Table 6.3: Clinical metadata of the Azizi dataset. Adapted from Azizi et al. 2018.

Patient	Tumour	Normal	Blood	Lymph node	Size (cm)	Metastases	Grade	ER	PR	HER2	Post-menopause	Age	Subtype	BRCA Deficiency
BC1	True	True	True	False	1	0	1	0.95	0.95	-	-	38	Ductal	Negative
BC2	True	True	False	True	3	1	2	0.9	0.1	-	+	60	Ductal	Unknown
BC3	True	True	False	False	1.5	0	3	0	0	-	-	43	Ductal	Negative
BC4	True	False	True	False	2.1	1	1	0.95	0.95	-	-	52	Ductal	Negative
BC5	True	False	False	False	2	0	3	0.05	0.01	-	+	78	Ductal	Unknown
BC6	True	False	False	False	1.3	0	2	0.99	0.01	-	+	58	Ductal	Unknown
BC7	True	False	False	False	1.2	0	3	0	0	+	+	65	Ductal	Unknown
BC8	True	False	False	False	1.3	0	2	0.2	0.05	-	+	72	Ductal	Unknown

The **TCGA dataset** consisted of bulk RNA-seq of breast tumours, as well as clinical data for 878 samples. These data were obtained from the data portal of TCGA, a project aimed at cataloguing cancer genomic profiles (Tomczak et al. 2015; Weinstein et al. 2013). Age information was only available for 756 samples. Count matrices were normalized in-house using the *voom* function from the limma package (Ritchie et al., 2015).

6.2.2. Normalization

Normalization of the Azizi dataset was performed using the *computeSumFactors* function, available at the scran package (Lun, Bach and Marioni 2016). The following analysis was performed using the Seurat package (Stuart et al. 2019), which is optimized for analysis of large scRNA-seq datasets.

6.2.3. Feature selection and data scaling

Identification of HVG was performed using the *FindVariableFeatures* function, with default parameters (Stuart et al. 2019). This method returned 2,000 features ranked by standardized variance, which represents a measure of single cell dispersion after controlling for mean expression (Stuart et al. 2019). In order to avoid dominance of highly-expressed genes in downstream analysis, we used function *ScaleData* (Stuart et al. 2019) to shift and scale the expression of each gene across cells, resulting in a mean equal to 0 and variance equal to 1.

6.2.4. Clustering

Clustering of the Azizi dataset consisted of sequentially applying three Seurat functions, namely *runPCA*, *FindNeighbors* and *FindClusters* (Stuart et al. 2019). *runPCA* performs linear dimensional reduction on the dataset. The obtained principal components scores are used as input by the *FindNeighbors* function, with each of the 10 first PCs representing a combination of correlated features.

This method implements a KNN (K-nearest neighbor) graph-based clustering approach, using the Euclidean distance in PCA space. Finally, clustering of cells is performed using the *FindClusters* function, which applies a modularity optimization technique, i.e., a measure of the strength of division of graphs into modules, to group cells together (Stuart et al. 2019). In order to use this function, we chose the resolution parameter, i.e., the level of detail of the clustering analysis, considering that the higher the resolution parameter, the higher the number of clusters obtained. Thus, we chose an intermediate value of 0.5 for the whole tumour-infiltrating immune cell dataset, which returned 17 clusters, an approximate number of the 19 originally labelled cell types and subtypes. For the clustering of the macrophage subset, which consisted of less than 3,000 cells, we expected two major polarizations. Hence, to avoid over-clustering, we used a lower value of resolution, 0.1, returning 3 clusters.

6.2.5. Marker genes

To identify marker genes for each cluster of the macrophage subset, we used the *FindMarkers* function (Stuart et al. 2019), comparing each cluster against the other two (cluster 0 vs. clusters 1 and 2, and so on). Using the *test.use* parameter, we selected the *MAST* (Model-based Analysis of Single-cell Transcriptomics) method to perform differential expression, which in turn utilizes the MAST package (Finak et al. 2015) to run the differential expression analysis. MAST uses a generalized linear model that accounts for the typical bimodal data of scRNA-seq datasets (with either strongly non-zero or non-detectable expression) and stochastic dropouts, which result in sparse count matrices (Finak et al. 2015).

6.2.6. Estimation of relative immune cell type proportions

To infer immune cell type relative abundance of bulk RNA-seq breast tumours, we used CIBERSORT. This step enabled the use of a microarray-derived gene signature, the LM22 (Chen et al. 2018), to perform immune cell type deconvolution of the TCGA breast tumour datasets, while minimizing batch effects as a source of confounding technical variation. LM22 is a gene signature consisting of 547 genes that distinguish 22 human hematopoietic cell subsets, including different subsets of T cells, macrophages, B cells, dendritic cells, mast cells, eosinophils and neutrophils (Chen et al. 2018).

6.2.7. Survival analysis

TCGA samples were divided based on estimated proportions of CD8+ T cells and M2 macrophages. The significance of differences in prognostic was estimated using Kaplan−Meier plots and log-rank tests, using R package survival (Therneau, 2015).

Kaplan-Meier estimates, applied in survival analysis, consist of measuring the number of subjects who survived over a period of time, starting from a defined point, such as cancer diagnosis, until the occurrence of the event, i.e., death. The Kaplan-Meier survival curve is defined as the probability of surviving in a given length of time. This method also allows not to exclude censored events, for example, when some of the subjects may not experience death before the end of the study, by considering that these patients have the same survival prospects as those who continue to be followed (Kishore, Goel and Khanna, 2010).

6.2.8. Differential gene expression analysis

We selected two groups of TCGA breast tumour samples for differential expression analysis: those with relatively high proportion of CD8+ T cells and low proportion M2 macrophages, and vice-versa. The criteria for the selection of each group was based on the results of the immune cell type deconvolution analysis. For the first group, we used a cut-off of a relative proportion of CD8+ T cells > 0.15 and a relative proportion of M2 macrophages < 0.12, selecting 7 samples. For the second group, we used a cut-off of a relative proportion of M2 macrophages > 0.44 and a relative proportion of CD8+ T cells < 0.05, selecting 7 samples (Supplementary figure 10.1).

Differential expression analysis was performed using the limma package (Ritchie et al., 2015), following three main steps. These included creating a design matrix distinguishing the previously selected groups by using R function *model.matrix*, fitting the model to obtain the pooled variance, using the *lmFit* function, and computing the moderated contrast t-test and B statistic (the log-odds that a gene is differentially expressed), using the *eBayes* function (Ritchie et al., 2015)..

6.2.9. Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005) was performed using the desktop application. We used a previously ranked list of genes, ordered by decreasing moderated t-statistic, derived from the differential expression analysis of breast tumour samples with relatively high and low content in CD8+ T cells and M2 macrophages. We excluded any differentially expressed genes from the list that were also part of the LM22 gene signature, to avoid possible biases resultant of the immune cell type deconvolution analysis. We ran the analysis using the Molecular Signatures Database (MSigDB) hallmark gene set collection (Liberzon et al. 2015). Gene sets with a with FDR < 0.1 were considered significant.

7. Data Analysis

7.1. Obtaining the gene expression signatures of the major brain cell types

7.1.1. Quality control and normalization

scRNA-seq can be affected by multiple technical artefacts arising from cell sorting, library preparation, and sequencing. Quality control to discard poor quality cells is therefore essential for reliable downstream analysis, ensuring that the biological signal of interest is not obscured by technical effects. Library size and the number of unique features detected are two quality metrics that can be used to find outliers.

The Spaethling dataset is a high-depth scRNA-seq dataset, with a median of 12,385,422 counts and 16,084.5 unique features detected per cell. If the detection rates were equal for all cells, we would expect to see an approximately normal distribution for these measures. Plotting the histograms for these measures revealed a left tail of cells with relatively small library sizes and few expressed genes (Figure

7.1). This is likely associated with failure to capture and convert RNA into cDNA during library preparation, hence, we discarded samples with library size lower than 1,000,000 counts and number of unique features detected inferior to 5,000.

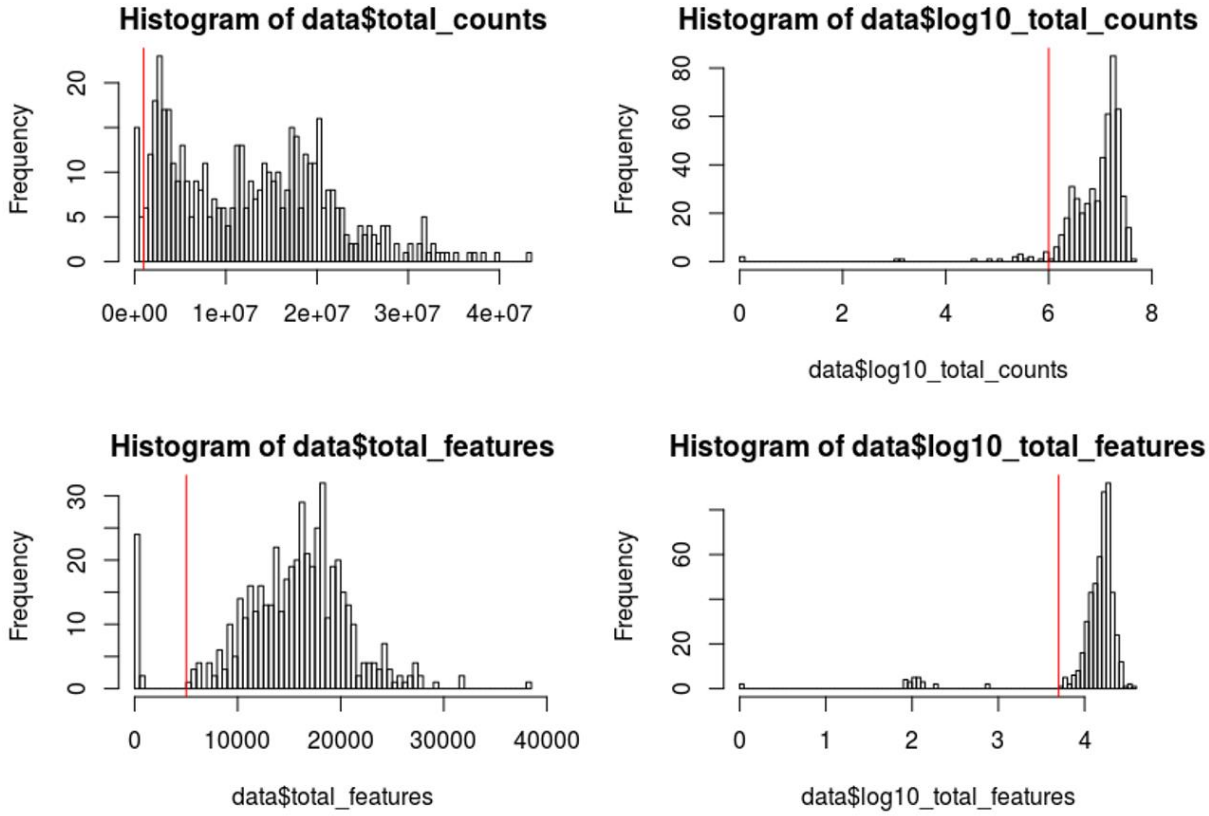

Figure 7.1: Histograms of the library sizes (total_counts, top panels) and the number of unique features detected (total_features, bottom panels) in all cells of the Spaethling dataset. The filter cut-offs are indicated by the red lines.

The percentage of mitochondrial genes is also an important quality metric to consider. Lysed cells lose cytoplasmic RNA, while the mitochondrial transcripts remain within the intact mitochondria. Apoptotic cells express mitochondrial genes and export these transcripts to the cytoplasm. Cells that show an increased expression of mitochondrial genes likely represent a group of dying cells. In this dataset, we filtered cells that showed over 25% of mitochondrial genes (Figure 7.2A). The original dataset also included mitochondria, which were used as positive controls, having 100% of mitochondrial genes (leftmost and uppermost point/cell in Figure 7.2A).

Another alternative to perform quality control consists of using the scater package to conduct PCA on a set of quality control metrics, which automatically identifies outliers based on the PCA components. Here, we used the number of unique features detected, library size, and the percentage of counts represented by the top 100 features (Figure 7.2B). Using the previous manual cut-offs, we discarded more cells than the automatic function of scater (Figure 7.2C), although all of the outliers captured by this function were also considered as such by the former manual method (Figure 7.2D).

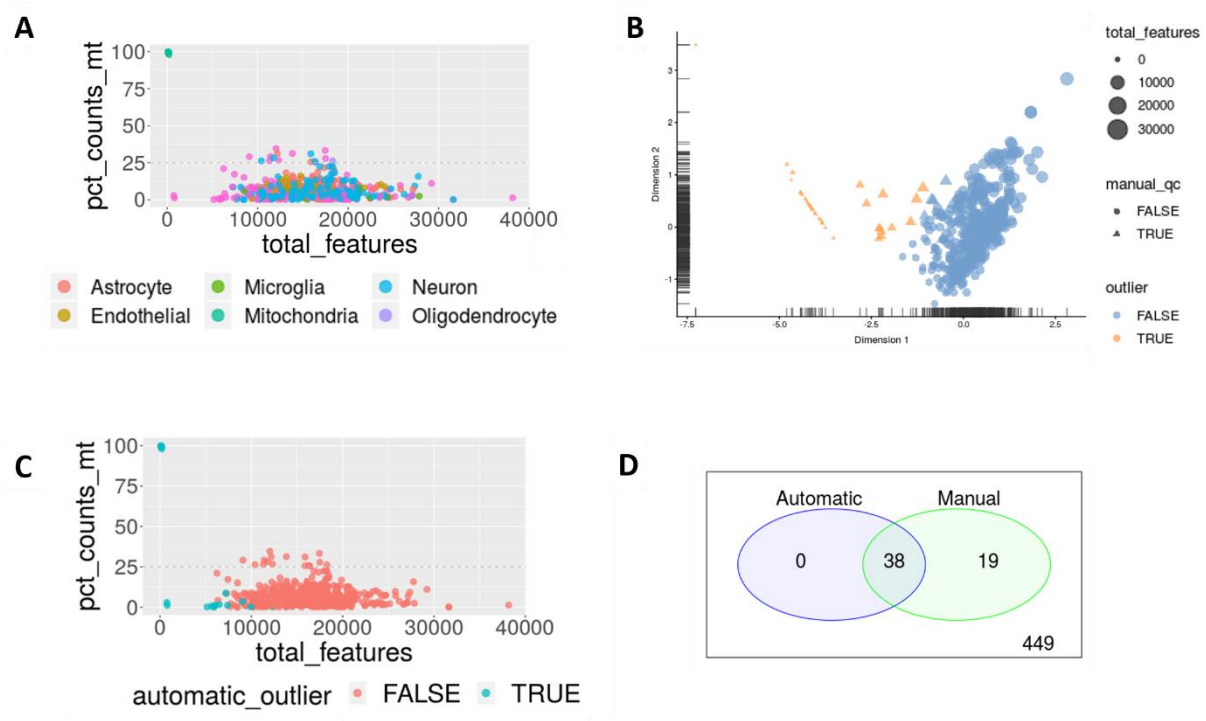

Figure 7.2: Quality control of the Spaethling dataset. ***A*** *- Scatterplot of the percentage of mitochondrial genes vs the number of unique genes, coloured by cell type;* ***B*** *– PCA generated by the automatic outlier detecting function from the Scater package, coloured by the outliers detected using manual cut-offs and shaped by whether cells or considered to be outliers by the automatic function;* ***C*** *- Scatterplot of the percentage of mitochondrial genes vs the number of unique genes, coloured by the outliers detected or not in the Scater function;* ***D*** *– Venn diagram of the outliers detected by the Scater function and by the manual cut-offs.*

After filtering low quality cells, we looked at the identities of the most highly expressed genes in the dataset (Figure 7.3). These are mostly constitutively expressed genes that encode for ribosomal and mitochondrial proteins. In the case of absence of ribosomal proteins or the presence of their pseudogenes, there could have been suboptimal alignment of the reads.

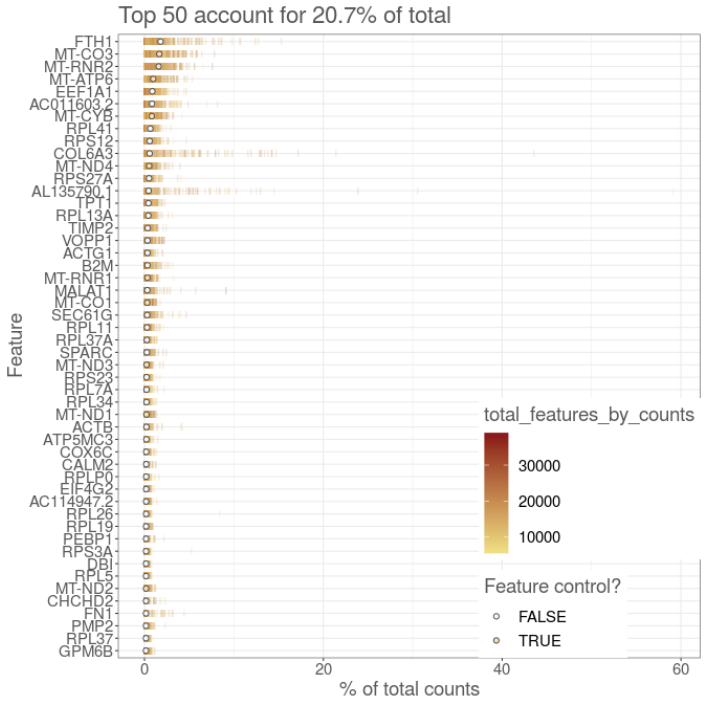

Figure 7.3: Percentage of total counts attributed to the top 50 most highly-expressed features in the Spaethling dataset. Each bar represents the percentage attributed to a feature for a cell, with the circle representing the average across all cells. Bars are coloured by the total number of expressed features in each cell and circles by whether the feature corresponds to a mitochondrial gene.

To facilitate the downstream analysis by reducing noise and the number of features, we filtered lowly expressed features by considering a threshold of at least five reads in at least two cells (Figure 7.4), which resulted in 12,216 features excluded, out of 57,799.

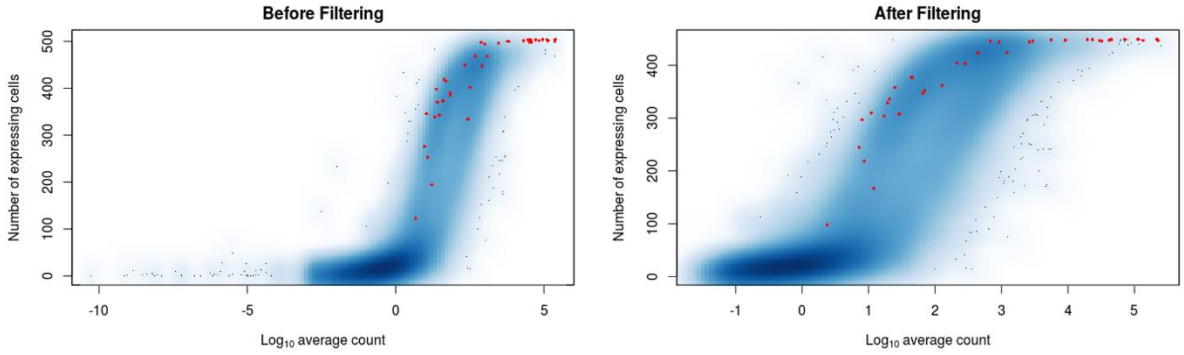

Figure 7.4: Number of expressing cells vs the log10-transformed mean expression for each feature in the Spaethling dataset, before (left) and after (right) filtering lowly expressed genes. Mitochondrial genes are highlighted in red.

In order to visualize the dataset, we applied dimensionality reduction techniques, namely, PCA and t-SNE, using the log-transformed count matrix (Figure 7.5A). In the PCA, the first component separated astrocytes and endothelial cells from microglia. Neurons and unidentified cells were scattered along this component, while the second component separated mostly astrocytes, neurons, and unidentified cells from microglia and endothelial cells. In the t-SNE, there were already clusters of cells being grouped by

cell type, although some neurons and unidentified cells remained relatively dispersed amongst other cell types.

To identify possible confounding variables, we plotted the percentage of the variance of expression values that is explained by the set of known variables (7.5B). The percentage explained by the number of unique features detected was relatively low, approaching 1%. Library size had a bimodal distribution, with the higher peak at 1% and the lower one at 10%. Sequencing date accounted for about 10% of variance in the dataset. Given that this variable represents mostly technical artefacts, it was considered a target for batch effect normalization. However, there were 52 different sequencing dates from the year 2013 to 2017, with most of the cells being sequenced in 2015. Therefore, we defined a new variable, *date_groups*, that consisted of joining the dates that comprised the minority of cells from different years into only one category.

Subject and body site were confounded with cell type, knowing that some cell types belonged mostly to one subject, and most subjects had tissue extracted from only one part of the brain. The remaining variables did not appear to account for much variance in the dataset.

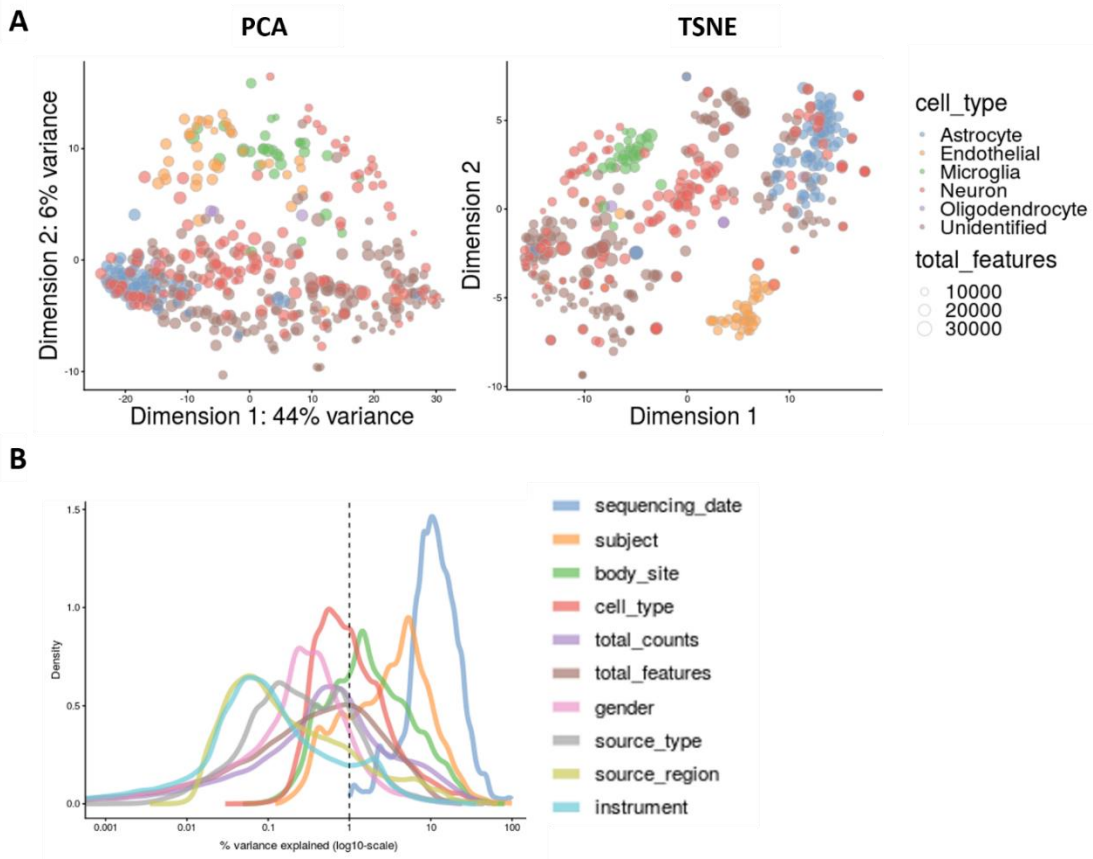

*Figure 7.5: **A**- PCA and t-SNE of the Spaethling dataset; **B** – Explanatory variables.*

After normalizing the dataset for library size and batch effect, the variance explained by variable *date_groups* notably decreased (Figures 7.6C and 7.6F). However, when looking at the t-SNE (Figures 7.6A, 7.6B, 7.6D and 7.6E), there still appeared to be clustering of cells by proximate dates, e.g. neurons sequenced on 08/2015 clustering with astrocytes sequenced on 09/2015. To check if normalization for batch effect was effectively removing the technical component of gene expression, we performed

pairwise differential expression analysis between different cell types, using the raw count matrix (before normalization), and the log-transformed count matrix (after normalization for library size and batch effect). This analysis revealed that normalization introduced a bias where genes with higher expression showed a tendency to have a higher moderated T-statistic (Figure 7.7B). In addition, batch effect correction led to a group of genes with very low expression exhibiting aberrant values of moderated T-statistic. Thus, we decided to continue the analysis without performing batch effect correction, given that it did not result in any advantage for this dataset, and tried to understand what were the drivers of proximity in the clustering analysis.

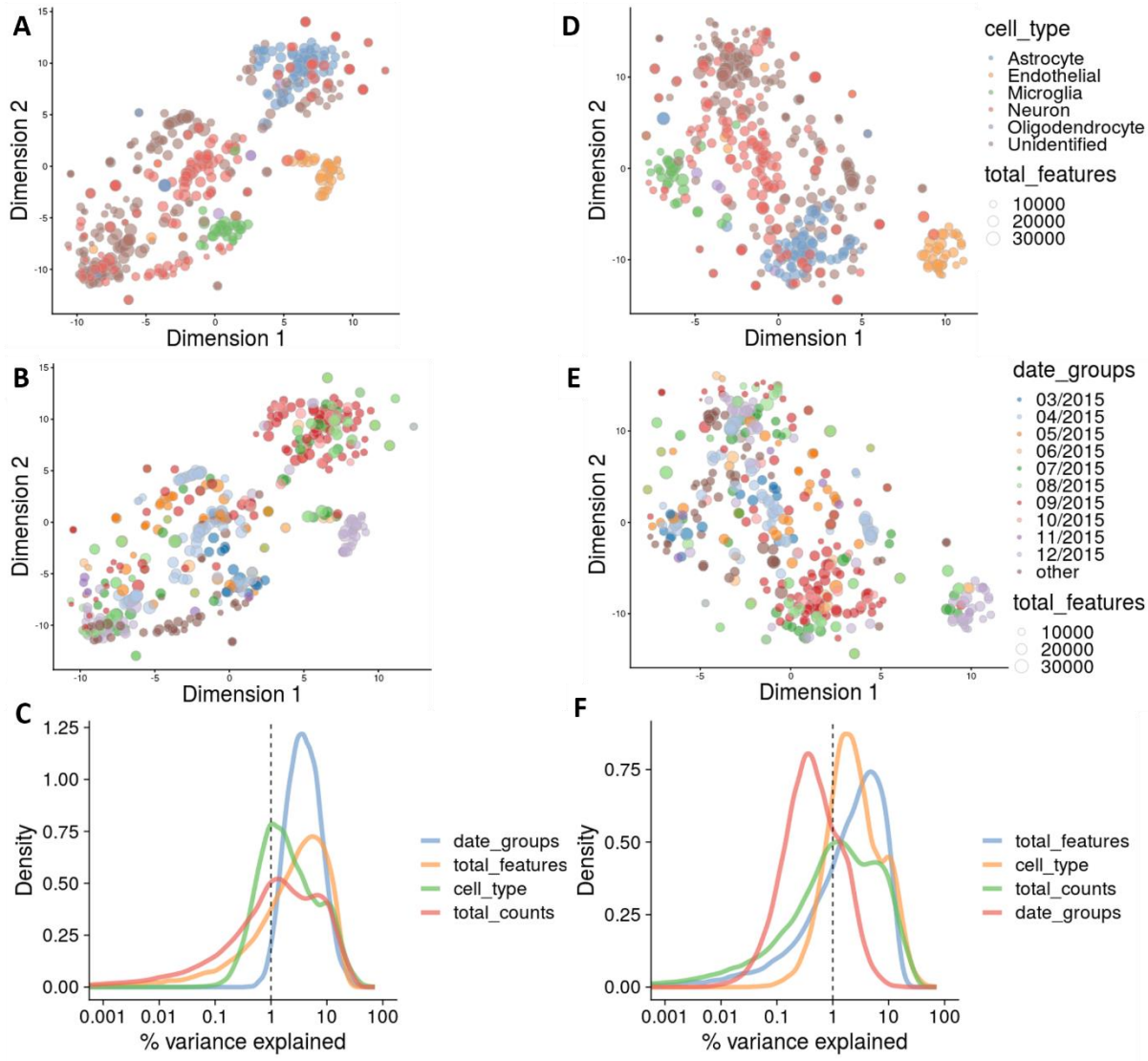

*Figure 7.6: Normalization and batch effect correction of the Spaethling dataset. **A** – t-SNE after normalization for library size, coloured by cell type; **B** – t-SNE after normalization for library size, coloured by the variable date_groups; **C** – Explanatory variables after normalization for library size; **D** – t-SNE after normalization for library size and correction of batch effect, coloured by cell type; **E** – t-SNE after normalization for library size and correction of batch effect, coloured by the variable date_groups; **F** - Explanatory variables after normalization for library size and batch effect correction;*

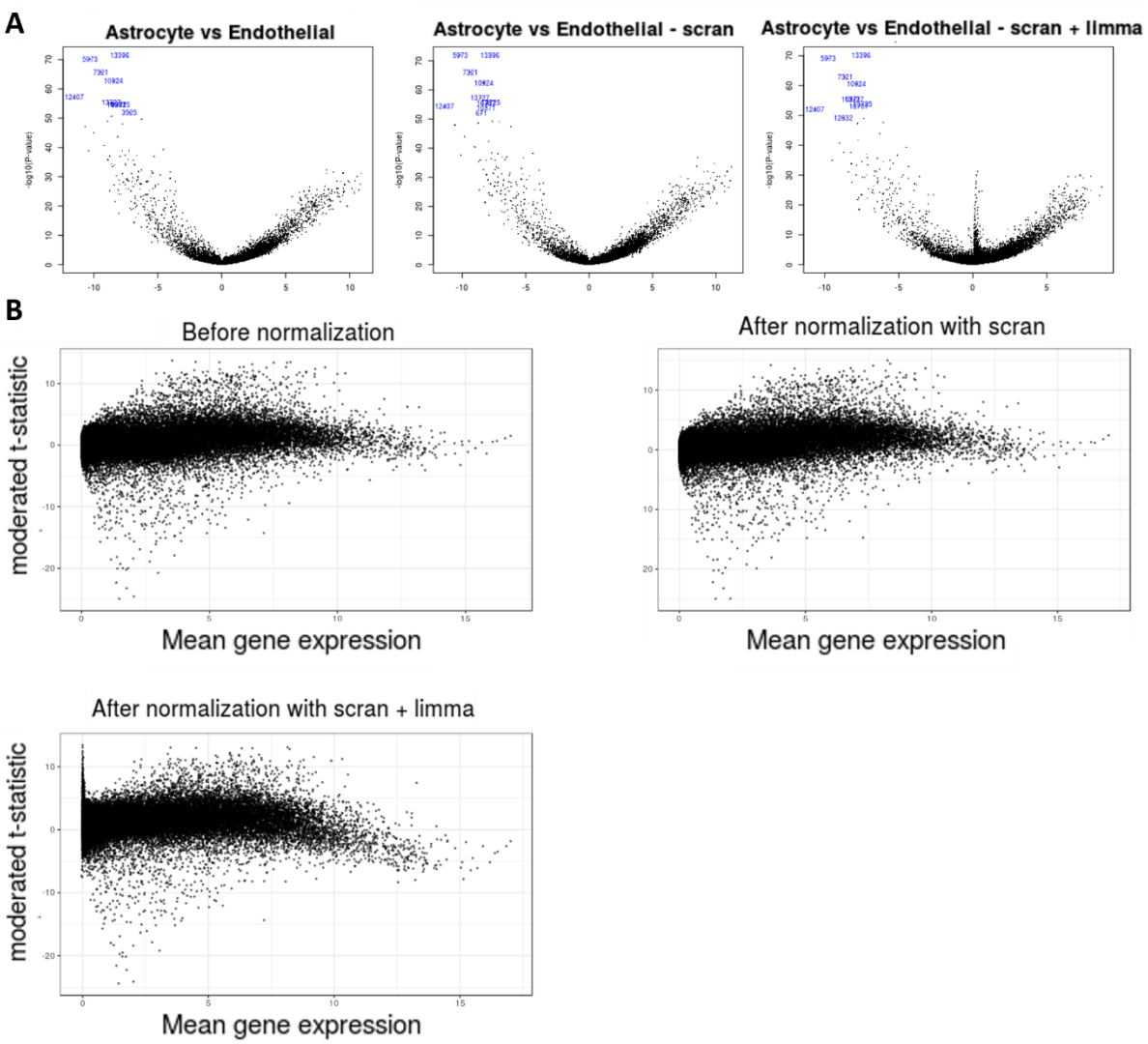

Figure 7.7:Example of the results of performing pairwise differential expression analysis in the Spaethling dataset, without normalization, after normalization for library size (scran) and after normalization for library size and correction of batch effect (scran + limma). ***A*** *– Volcano plots resultant of the comparison of astrocytes and endothelial cells;* ***B*** *– Scatter plot of moderated t-statistic vs mean gene expression for the astrocyte and endothelial cell comparison.*

7.1.2. Pseudotime analysis

The t-SNE of the Spaethling dataset (Figures 7.6A, 7.6B, 7.6D and 7.6E) resembled a trajectory between unidentified cells, neurons and astrocytes (Supplementary figure 10.2). Given that cells sequenced in this dataset were submitted to long-term cell culture, this could mean that we were dealing with different stages of differentiation. Thus, we conducted a pseudotime analysis to check for an underlying developmental trajectory. The results indicate that there was indeed a trajectory, with a root composed of astrocytes, neurons and unidentified cells, followed by unidentified cells with intermediate pseudotime values and culminating in a cluster of mostly neurons and unidentified cells (Figure 7.8A). An alternative state of differentiation can be suggested, represented by microglia and endothelial cells. When plotting the t-SNE coloured by pseudotime score, we see a clear correspondence between the cells position and their pseudotime score (Figure 7.8B).

These results are in concordance with the original publication (Spaethling et al., 2017), in which astrocytes were described as being actively dividing, which meant they were less differentiated than the remaining cell types. Unidentified cells were not assigned a cell type due to having ambiguous morphology traits and marker genes that could correspond to either neuronal or glial cell types. We speculate these cells could also represent an undifferentiated cell type, such as radial glia, capable of generating both neurons and astrocytes (Barry, Pakan and McDermott 2014). It is also known that astrocytes are able to convert into neurons under specific conditions, so we can speculate that in a hostile environment, resembling a wounded brain, astrocytes may suffer a transition and develop into cells that have properties both similar to astrocytes and neurons (Cheng et al. 2015; Laywell et al. 2005).

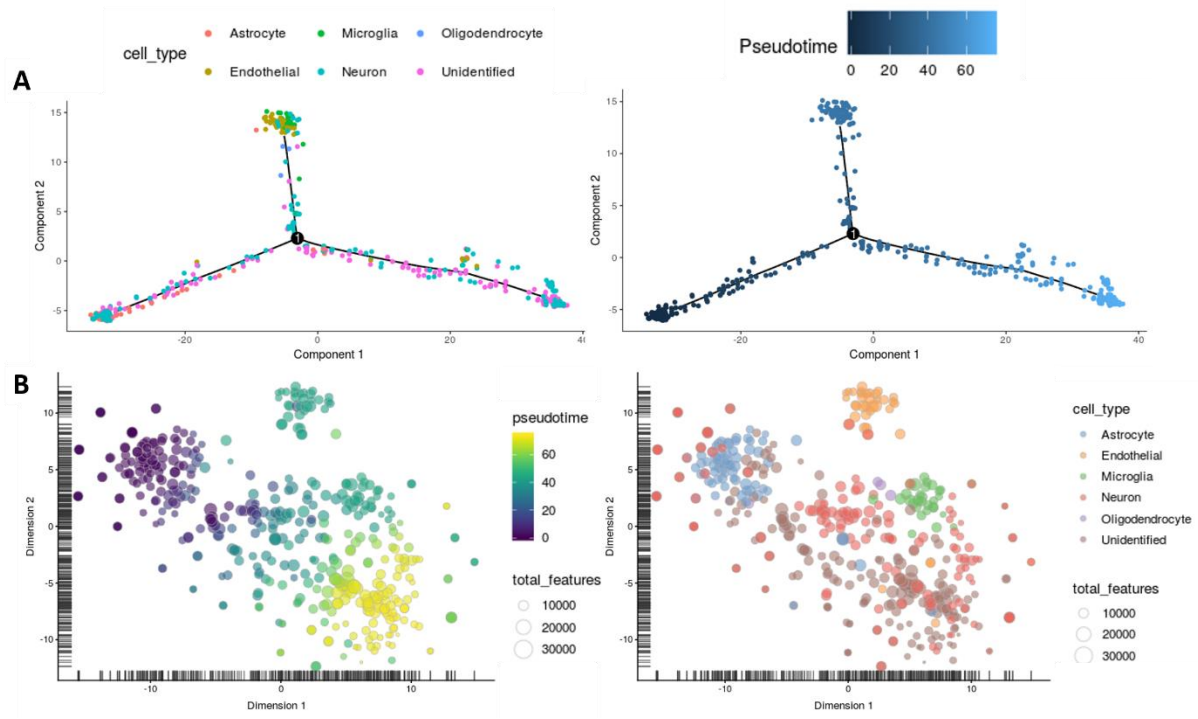

*Figure 7.8: Pseudotime analysis of the Spaethling dataset. **A** – Visualization of the trajectory in the reduced dimensional space, coloured by cell type (left) and by pseudotime (right); **B** – t-SNE coloured by pseudotime (left) and by cell type (right).*

7.1.3. Feature selection and clustering

In most scRNA-seq experiments, only a portion of genes show a response to the biological condition of interest, e.g., differences in cell-type. Feature selection is a useful step to remove genes which were only detected at different levels due to technical noise. This step is also important in increasing the computational efficiency of the analysis, by reducing the total amount of data necessary to process, e.g., in clustering methods.

In this project, we tested two different methods, which consisted in the identification of highly variable genes (HVG) and high dropout genes (HDG). Finding HVG consists on assuming that some genes have large differences in expression across cells due to biological variability, rather than technical noise, after correcting for the positive relationship between variance and mean expression. Finding HDG consists of identifying genes with unexpectedly high dropout rates. Dropout rates, as discussed in section 4.2, are related with average expression level, and result from a failure in mRNA reverse transcription. Because reverse transcription is an enzymatic reaction, it can be modelled using the Michaelis-Menten

equation. Using the HVG and HDG methods, we identified 2,351 and 6,932 features, respectively, with 677 significant features in common between both methods (Fisher's Exact Test, p-value < 2.2e-16).

Next, we performed consensus clustering to verify if cells clustered by cell type. We tested the use of all of the features captured in the dataset, the features selected with HDG, the features selected with HVG, and the features selected in common with both methods (Figure 7.9). We found that cells tended to cluster according to their cell type, although the astrocyte and microglia clusters also contained neurons. Endothelial cells have a different origin relative to neuronal and glia cells, and, therefore, formed very clean clusters. Instead of clustering with cells for which their type was known, unidentified cells consistently formed separated clusters, except for a few that clustered together with astrocytes and neurons.

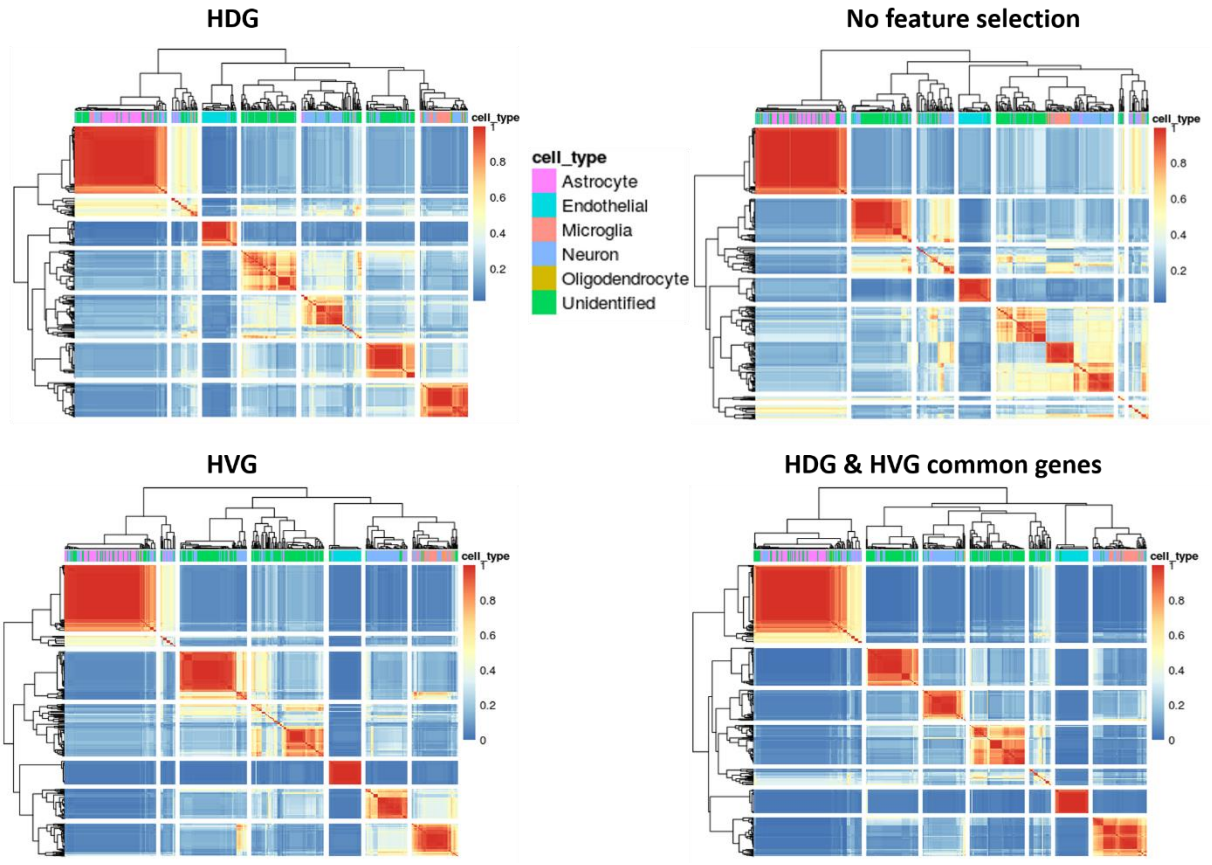

Figure 7.9:Consensus matrices of the Spaethling dataset, clustered based on log-transformed gene counts, using the whole set of available features (right top), using selected features with HDG or HVG (left top and bottom, respectively), and using common features between the latter methods.

These results were consistent with the pseudotime analysis, in which a group of unidentified cells appeared to be in the same state as astrocytes, and the rest followed a trajectory to the end of the branch (Figure 7.10).

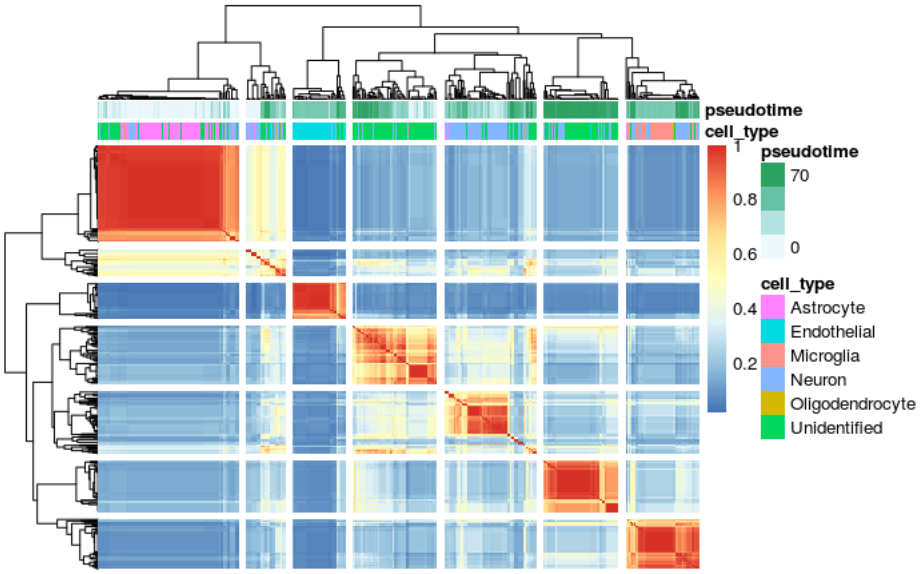

Figure 7.10:Consensus matrix of the Spaethling dataset, , clustered based on log-transformed gene counts , using only selected features with HDG. Pseudotime (state of differentiation) along with cell type seem to be the main drivers of cell similarity.

Carrying out feature selection proved to be beneficial for the clustering analysis, obtaining cleaner clusters of cell types comparatively to performing clustering with the full set of genes (Figure 7.9 and 7.10). The use of the common features, selected with both tested methods, generated the cleaner consensus matrix, with the majority of cells assigned to the same cluster after multiple clustering attempts using different parameters. This confirms the usefulness of feature selection methods in capturing relevant features, i.e., with variability in expression due to biological signal, thus reducing noise in the dataset.

Before determining the neuronal and glia gene signature, it was important to obtain very pure clusters of cells types, considering the uncertainty in the preliminary cell type labelling. Therefore, we used the consensus clustering results to remove ambiguous cells, i.e., cells present in a cluster where the majority of cells were of a different type. We also put aside unidentified cells, as the majority of their identities were not clear at this stage.

7.1.4. Marker genes

In order to find marker genes for each population of different cell types, we tested different combinations of feature selection and classification methods. As expected, the number of defined marker genes varied not only when using different algorithms of classification, but also when using the same algorithm for classification and varying the feature selection method (Table 7.1). The classification algorithm applied in the *SC3* function returned the lowest number of genes. However, the utilization of this function implies the application of user-selected cut-offs, for AUROC and p-value, which significantly affects the number of marker genes considered.

Table 7.1: Combination of feature selection and classification methods, to obtain the marker genes for each neural cell population. The combination of HDG & HVG common genes and SC3 returned under 50 marker genes. Therefore, it was excluded from the analysis.

Method	Feature selection	Marker genes	No. of marker genes obtained
1	HDG	PAMR	2582
2	HVG	PAMR	680
3	HDG & HVG common genes	PAMR	346
4	HDG	SC3	375
5	HVG	SC3	71
6	HDG	Elastic nets	588
7	HVG	Elastic nets	2344
8	HDG & HVG common genes	Elastic nets	287

We tested the resultant gene signatures with the cell type deconvolution tool, CIBERSORT, by using aggregated pseudobulk data, which consisted of summing count measurements from different proportions of cell types (Figure 7.11). In general, all of the combinations of methods generated gene signatures capable of estimating the gross relative abundances of each cell type. The gene signature from combination 1, in particular, demonstrated a relatively high level of concordance in this task, which was expected, given that it had the higher number of genes.

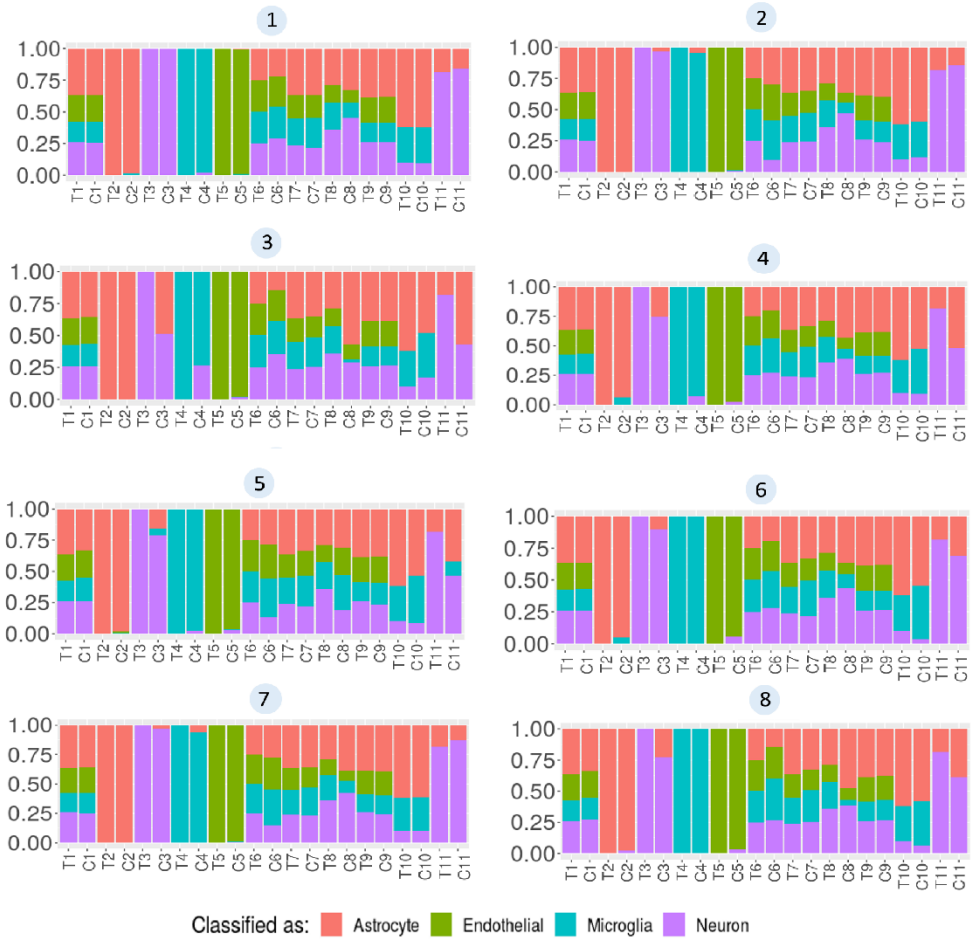

Figure 7.11: Results of performing cell type deconvolution on pseudobulk data, using gene signatures that resulted from each of the combinations of methods used. Each bar plot contains pairs of bars representing the expected scores for each cell type (T) and the scores obtained using the deconvolution tool (C). The combination, as in Table 7.1, used to obtain the gene signature is identified on top of each plot.

To ascertain whether these gene signatures were capable of generalizing the discrimination of the major brain cell types in external datasets, we tested them for their ability to correctly classify neurons, astrocytes and microglia cells from the scRNA-seq Darmanis dataset, using CIBERSORT.

Combination 1 generated the only gene signature capable of classifying neurons (Figure 7.12). However, this combination failed to classify astrocytes, as did the remaining combinations (Figure 7.13). Although combination 5 and 7 exhibited less noise in the classification of astrocytes comparatively to the remaining methods, they also incorrectly classified most neurons as being mostly astrocyte-like, which could indicate the existence of overfitting, i.e., the model was incorrectly optimized for astrocytes, resulting in the elevated number of false positives. Notably, combination 5 using the *SC3* function generated a gene signature with 71 genes that did not perform worse than the gene signature with 2344 genes obtained using combination 7 with elastic nets. Furthermore, microglia in general appeared to be an easier cell type to properly classify in an external dataset (Figure 7.14). These results were expected, given that microglia arise from a different progenitor than that of the remaining glia cell types, while astrocytes and neurons can share a common progenitor from which they differentiate. Hence, microglia cells likely have a gene expression profile more dissimilar to astrocytes and neurons.

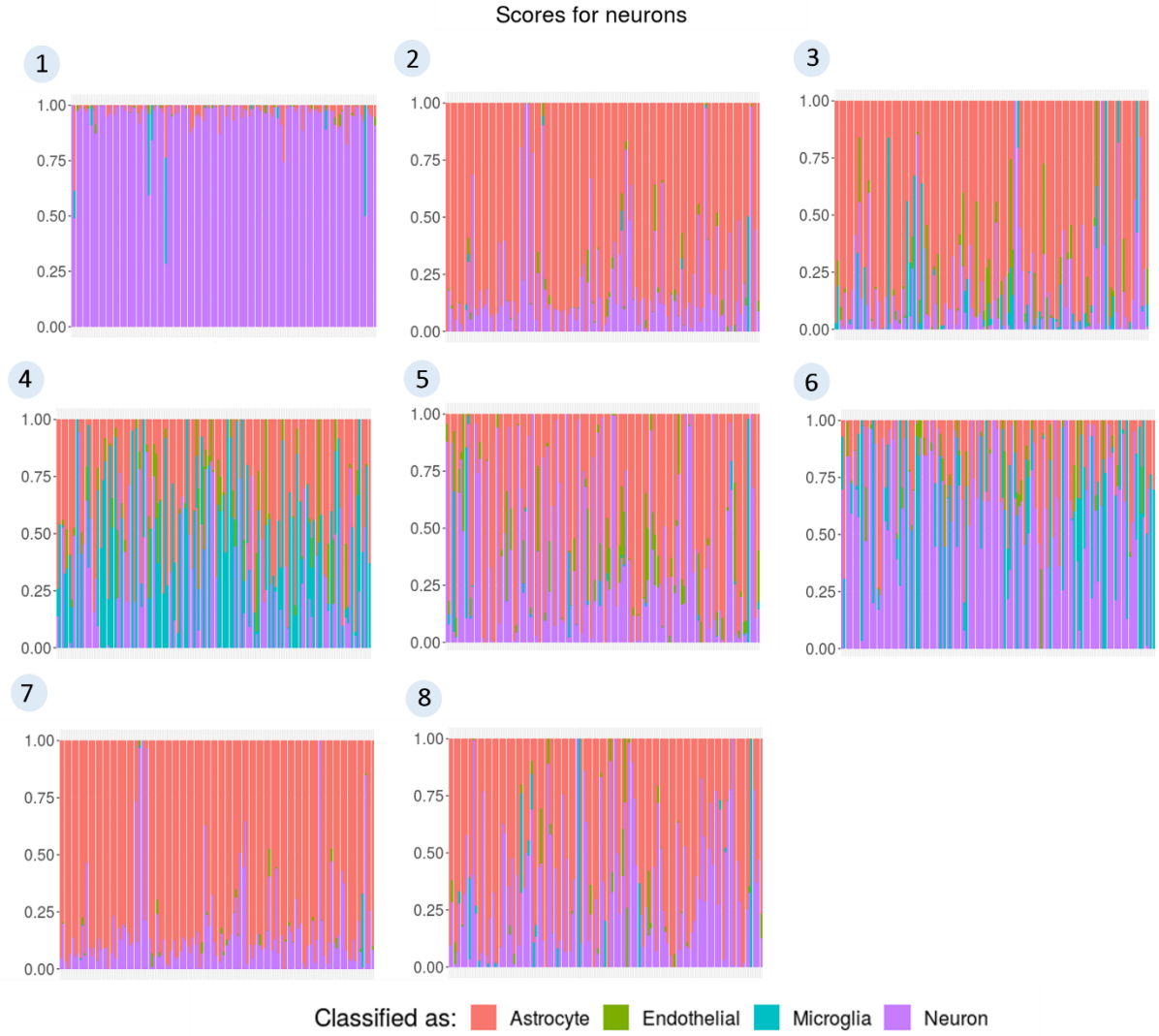

Figure 7.12: Results from performing cell type deconvolution with gene signatures generated by combinations of methods 1 to 8. This analysis was performed only on neurons from the Darmanis dataset.

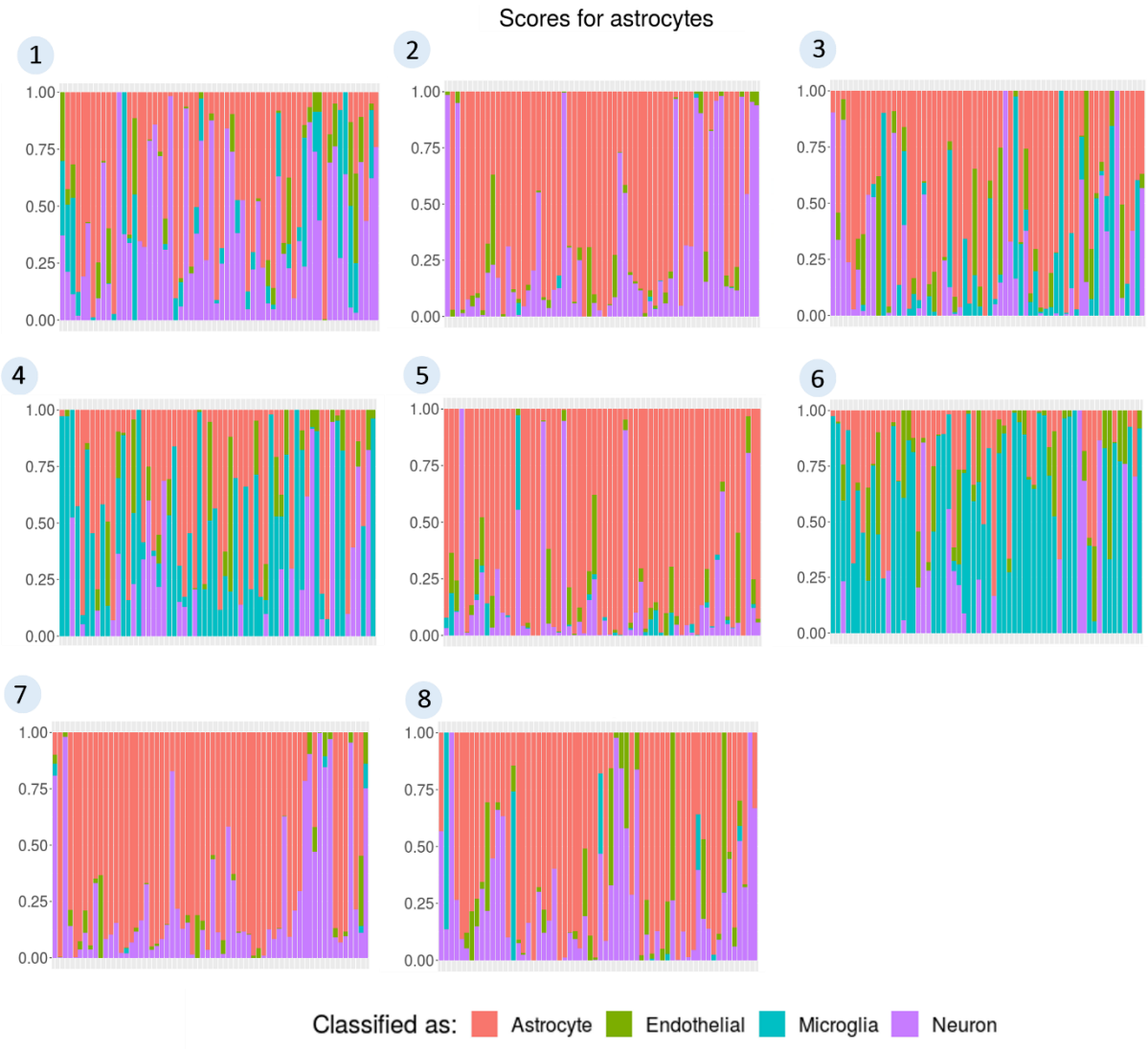

Figure 7.13:Results from performing cell type deconvolution with gene signatures generated by combinations of methods 1 to 8. This analysis was performed only on astrocytes from the Darmanis dataset.

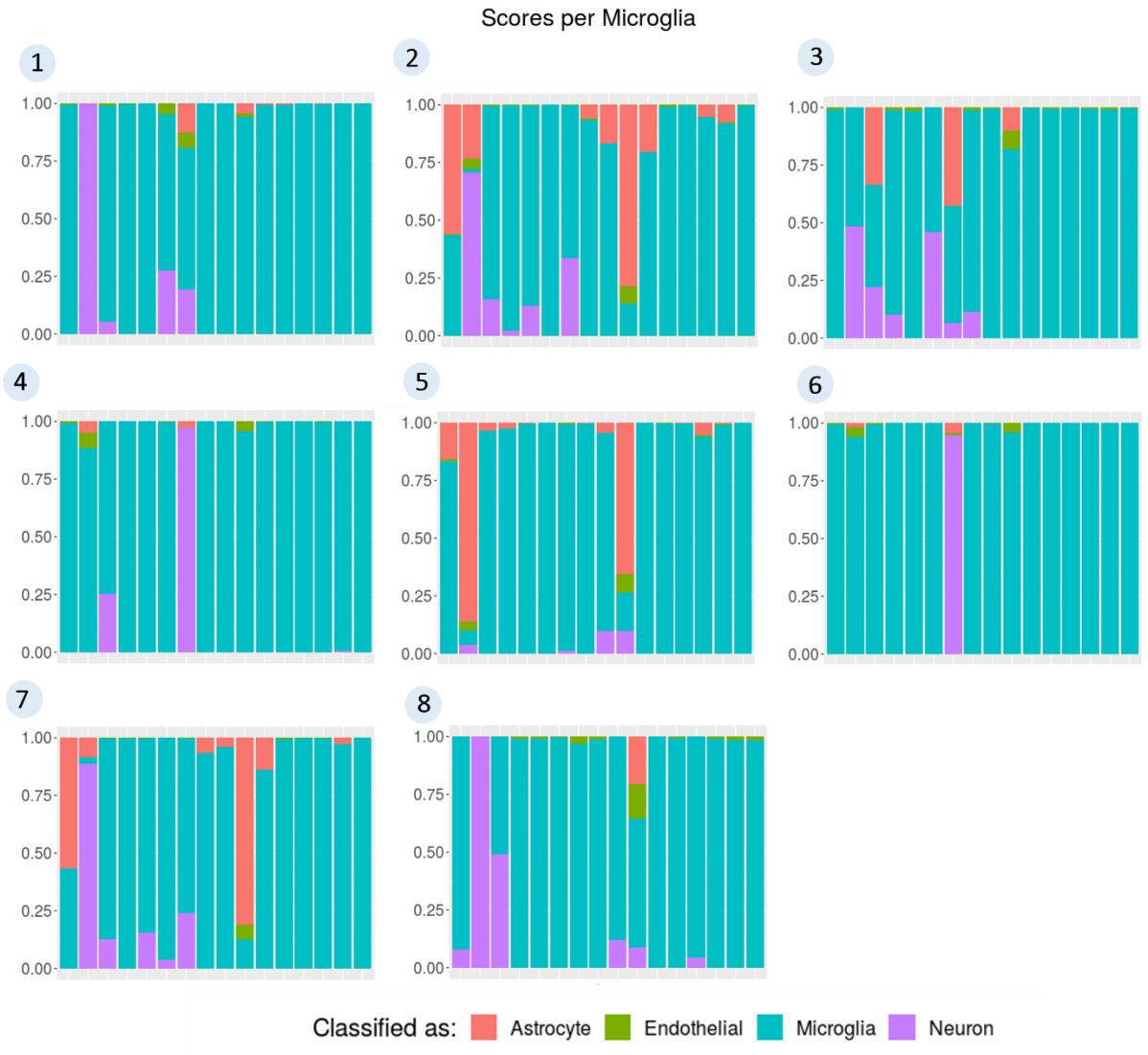

Figure 7.14: Results from performing cell type deconvolution with gene signatures generated by combinations of methods 1 to 8. This analysis was performed only on microglia from the Darmanis dataset.

Summarizing, the SC3 method is useful in reducing the number of marker genes, without extensively compromising the accuracy of cell type deconvolution in pseudobulk data. Considering that all the gene signatures under-performed in the generalization task with the external dataset, we hypothesized this was due to variability in gene expression resulting from one dataset having cells acutely removed from the brain environment (Darmanis) and the other cultured before sequencing (Spaethling). In order to further explore this hypothesis, and to obtain a gene signature that captured the common gene expression profile between cells of the same type and of different states, we decided to continue the analysis by merging the two datasets.

7.1.5. Merging scRNA-seq datasets

In order to merge the Spaethling and Darmanis datasets, we decided to exclude endothelial cells from the analysis, given that the latter dataset did not include this cell type and also taking into consideration that endothelial cells were expected to comprise only a very small proportion of brain tissue.

After combining both datasets, there was a clear grouping on the t-SNE based on the dataset of origin (Figures 7.15A - 7.15F). The Darmanis dataset also had smaller library sizes, which was expected given the different sorting and sequencing protocols. After normalization, the differences in library sizes were corrected but the cells remained clustered by dataset. Hence, we performed batch correction, after which cells clustered based on their cell type, with sub-clustering occurring to some extent between within cells of the same type and of the same dataset of origin.

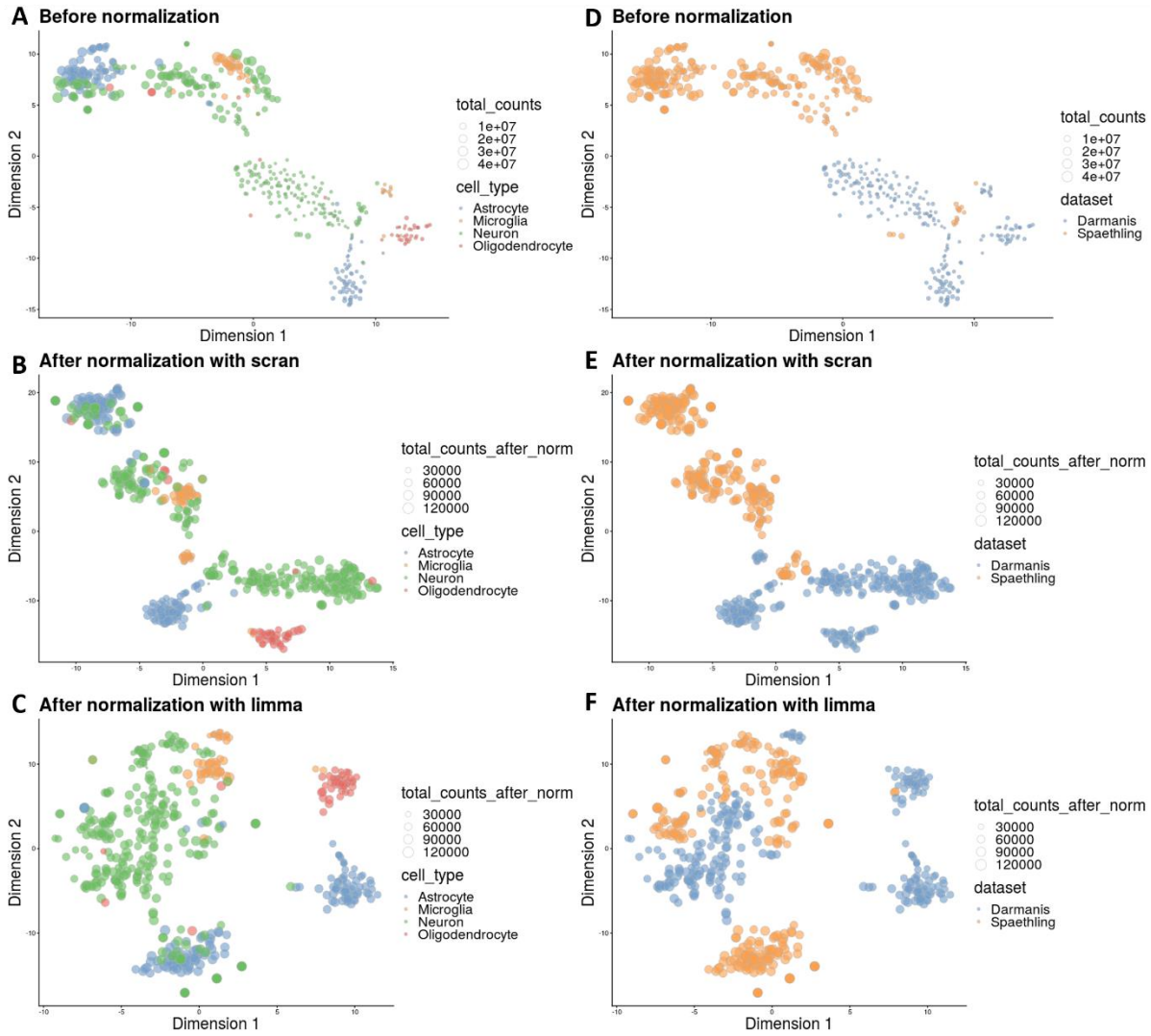

Figure 7.15: t-SNEs of the merged Spaethling and Darmanis dataset, before normalization (A,D), after normalization for library size (B, E), and after normalization for batch effect (C, F). Plots are coloured by cell type (left) and by original dataset (right).

Although this correction approximated astrocytes, they still formed two very distinct clusters based on the original dataset. We considered that this could be due to not only technical noise but also biological variability, since isolated cells from the Spaethling dataset were cultured up to 84 days *in vitro* before sequencing, and astrocytes were described as being actively dividing, contrasting with astrocytes from the Darmanis dataset, which were acutely isolated and sequenced. It is known that cultured astrocytes develop a reactive phenotype, as if their environment was disrupted in the case of brain lesion (Schiweck, Eickholt, and Murk 2018; Liddelow and Barres 2017). In addition, after normalizing for batch effect, the variance explained by the original dataset was significantly reduced, from reaching over 10% to about 0.01% (Figures 7.16A – 7.1C).

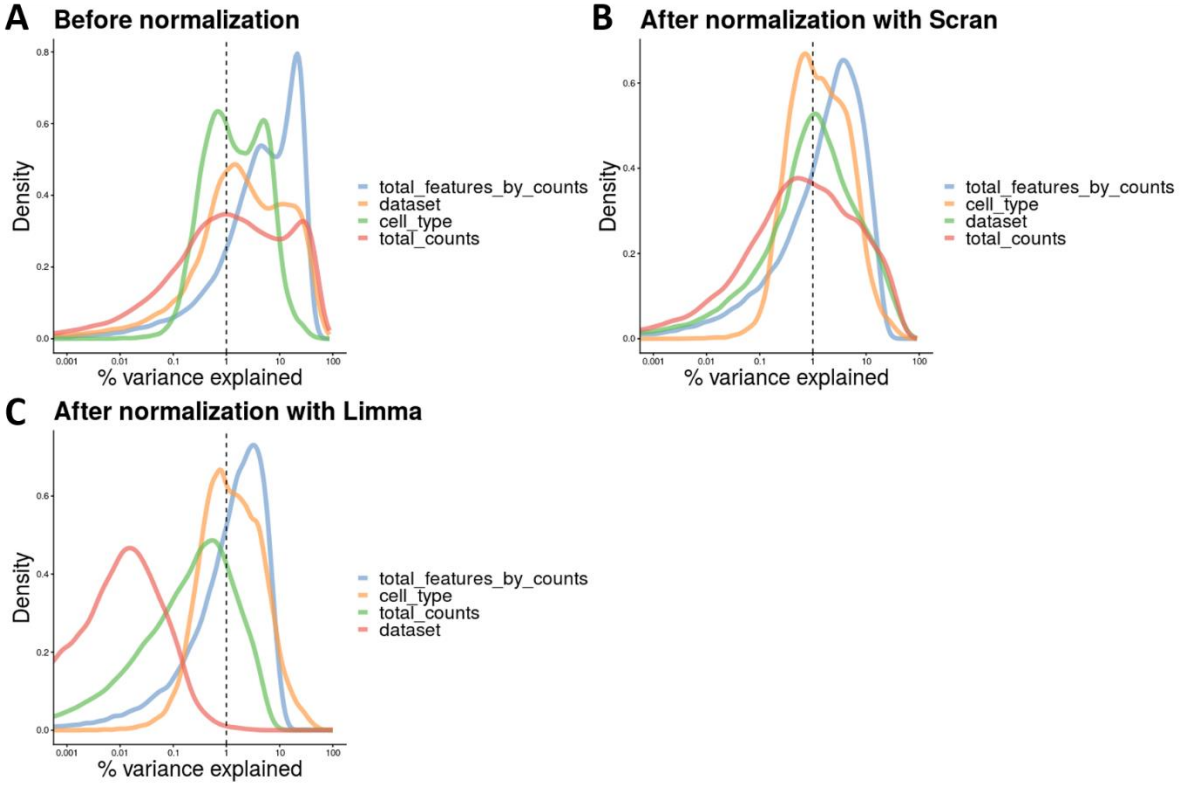

Figure 7.16: Explanatory variables of the Spaethling and Darmanis merged dataset, before normalization (A), after normalization for library size (B), and after normalization for batch effect (C). The plotted variables correspond to the number of unique genes detected (total_features_by_counts), library size (total_counts), cell type and dataset of origin.

To confirm that the existence of two different groups of astrocytes was due to biological variability, we used the *SC3* function to obtain the top 100 marker genes for each group of astrocytes and performed Gene Ontology (GO) enrichment analysis. As hypothesized, GO annotations for the Spaethling astrocyte marker genes were mostly associated with a reactive phenotype, e.g., astrocyte activation and neuroinflammatory response (Supplementary table 10.1). On the other hand, Darmanis astrocytes' annotations were related with the astrocytic role in controlling the levels and activity of several neurotransmitters, e.g., L-glutamate and D-aspartate import across plasma membrane (Supplementary table 10.2). Because these represent two different states of astrocytes, we decided to keep them as separate cell sub-types in the downstream analysis.

7.1.6. Gene signature for brain cell types

To obtain the final gene signature of the major cell types of the brain, we performed HDG feature selection on the merged dataset, which returned 7,662 selected features. Following, we tested both PAMR and the *SC3* function, resulting in 2321 and 1126 features, respectively. Validation of these signatures was performed using a train/test split of the dataset balanced by cell type and dataset of origin, i.e., we used the train dataset (266 cells) to obtain the marker genes, and validated the gene signature by performing cell type deconvolution on the test dataset (173 cells). Considering that both signatures were able to properly classify neurons, astrocytes, microglia and oligodendrocytes in the test dataset, we looked for marker genes common to both approaches, to diminish the number of necessary features. This returned our final gene signature, with 889 significant common marker genes (Fisher's Exact Test, p-value < 2.2e-16).

Finally, to confirm that the gene signature was able to be generalized to external datasets, we performed cell type deconvolution on the scRNA-seq Zhang dataset (Figure 7.17). We were able to properly classify human and mouse astrocytes, considering the astrocyte score as the sum of scores from different astrocyte states. Although there was some overlap between astrocyte marker genes of different states, astrocytes from the Zhang dataset were more similar to astrocytes from the Darmanis dataset, as expected, given that the latter were also acutely removed and sequenced.

Furthermore, neurons and microglia were also correctly classified. However, this method failed to correctly classify oligodendrocytes, with only 2 out of 5 showing a score over 0.5 for the oligodendrocyte cell type. This may be due to the fact that oligodendrocytes are difficult to isolate and may contain myelin and axon debris (Pfenninger et al. 2007; Tham et al. 2003).

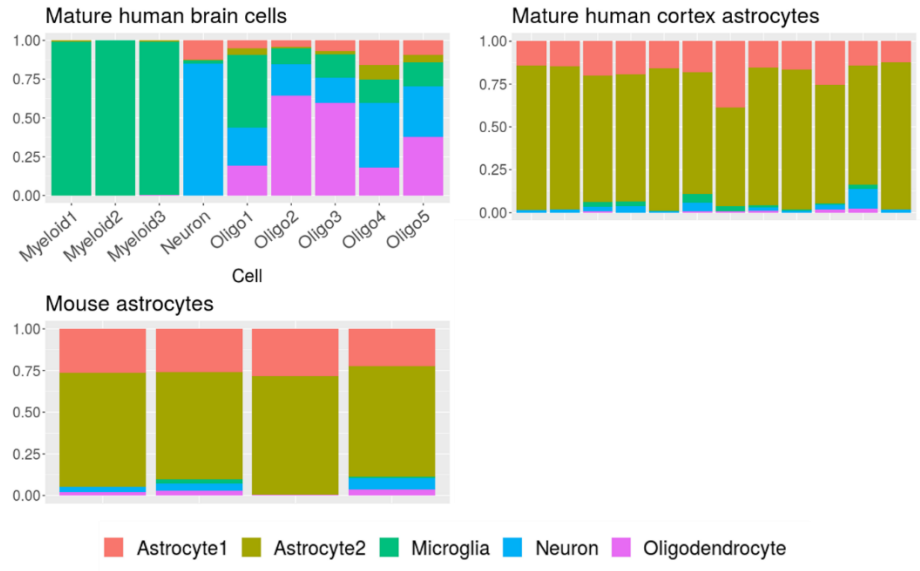

Figure 7.17:Results of cell type deconvolution on the Zhang dataset, with the final gene signature.

7.2. Unveiling how the relative abundance of neurons and glia in the brain correlates with ageing and neurological health

The following analysis consisted of using the previously defined and validated neuronal and glia gene signature to perform cell type deconvolution of publicly available bulk RNA-sequencing datasets of AD, PD and non-AD-PD brain tissues.

7.2.1. Cell type deconvolution of the healthy brain

We performed neural cell type deconvolution of bulk RNA-seq datasets of non-diseased brain tissue, available from GTEx. We started by looking at the whole set of brain tissues available, distributed among amygdala, anterior cingulate cortex, caudate (basal ganglia), cerebellar hemisphere, cerebellum, cortex, frontal cortex, hippocampus, hypothalamus, *nucleus accumbens* (basal ganglia), *putamen* (basal ganglia), spinal cord (cervical c-1), and *substantia nigra*. Next, we selected specific neural tissues to explore in further detail, namely cortex, frontal cortex, anterior cingulate cortex, cerebellar hemisphere, hippocampus, *substantia nigra*, and spinal cord.

When considering the analysis performed on the total set of available tissues, the most abundant cell type was neuron, followed by astrocyte, oligodendrocyte and microglia (Figure 7.18). These proportions are only representative of the amount of mRNA in each cell type. Currently, there is still no method to convert these values into true cell proportions. However, there is a correlation between the amount of mRNA and the number of a given cell type, which is useful for the analysis intended in this work.

Taking the above into consideration, we found that the relative proportions of neural cell types may vary extensively, depending on the region of the brain that is being analysed. For example, there were tissues with a relatively high median of the relative proportion of neurons, namely cortex (0.7234), cerebellar hemisphere (0.8723), and frontal cortex (0.7694), in contrast with *substantia nigra* (0.5195), hippocampus (0.6201) and spinal cord (0.2577). The proportion of astrocytes did not show as much variability as the proportion of neurons, albeit the cerebellar hemisphere showing a relatively lower median than the rest of the tissues (0.11812). Most tissues presented a very low proportion of oligodendrocytes, with exception of the spinal cord, followed by the hippocampus and *substantia nigra*, although the two latter had lower median values (0.41219, 0.09933 and 0.144677, respectively). The same occurred for microglia, with the spinal cord and *substantia nigra* presenting the higher median values (0.08074 and 0.01383, respectively), although the relative abundance of this glia cell type was, in general, very low in every tissue.

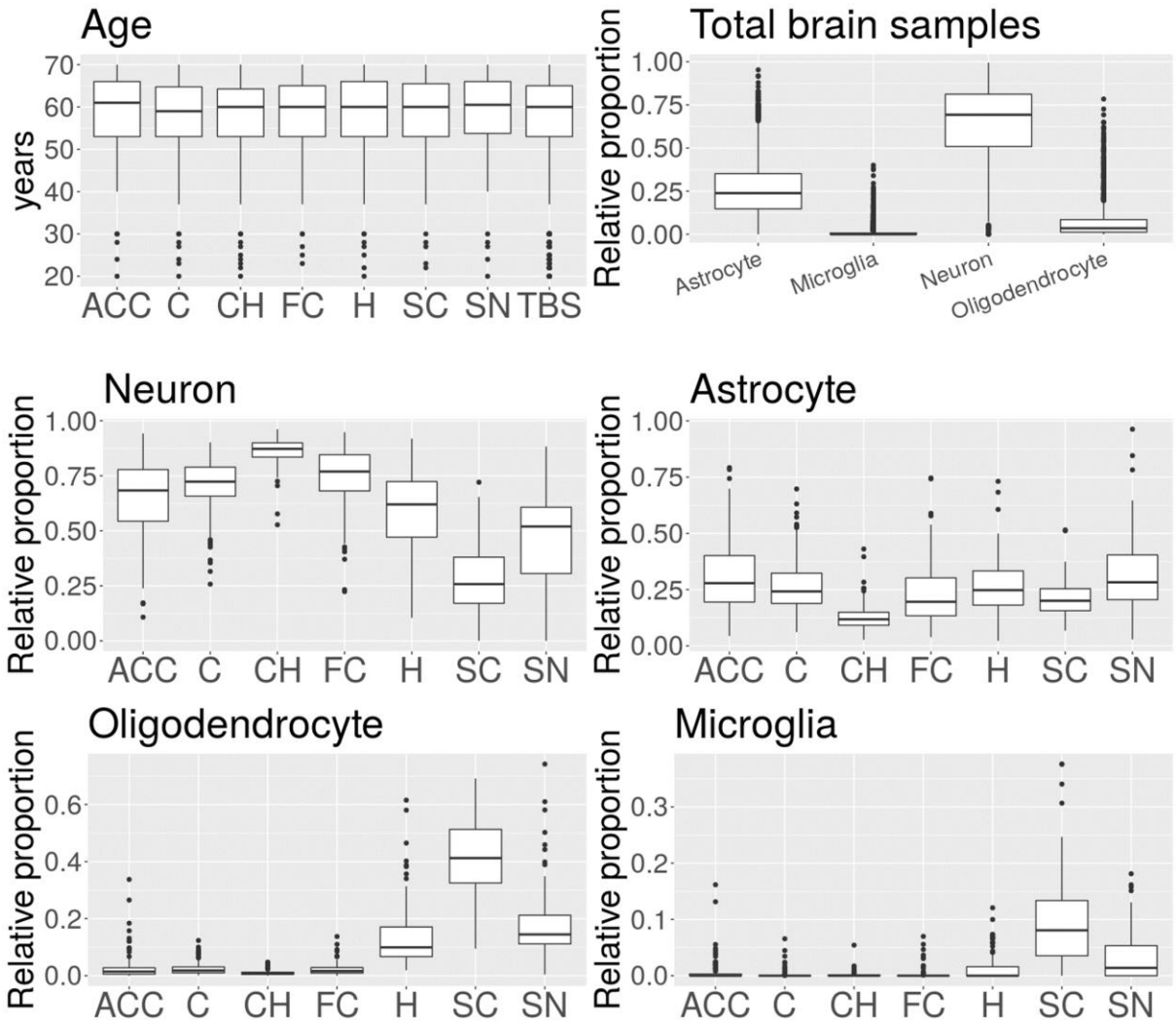

Figure 7.18: Box plots of age and relative proportions of brain cell types, grouped by brain tissue. TBS – Total brain samples; ACC – Anterior cingulate cortex; C – Cortex; CH – Cerebellar hemisphere; FC – Frontal cortex; H – Hippocampus; SC – Spinal cord; SN – Substantia nigra.

We further investigated whether there was a correlation between the relative proportion of each cell type and age, for each of the brain tissues selected (Table 7.2). With exception of the cerebellar hemisphere, every tissue showed a significant negative correlation between the relative proportion of neurons and age, and a concomitant increase in one or several types of glia cells.

Regarding the cerebellar hemisphere, as mentioned in the previous analysis, this tissue presented a higher relative proportion of neurons comparatively to other brain tissues, and, therefore, the variation of neuronal relative proportions with ageing in this tissue might not be enough to be captured using this method. However, it was also the only tissue that exhibited a negative correlation between the relative proportion of oligodendrocytes and ageing, which could be an indicator of the occurrence of demyelination with ageing in the cerebellar hemisphere (Adamo, 2013).

Table 7.2: Results of the Spearman's correlation analysis between the relative proportions of brain cell types in each analysed tissue and age. Green and red rows highlight significant ($p < 0.05$) and possible meaningful positive and negative correlations, respectively.

	Cell type	Correlation coefficient	P-value
Total brain samples	Neuron	0.089	0.00028
	Astrocyte	-0.065	0.0074
	Oligodendrocyte	-0.079	0.0012
	Microglia	0.015	0.55
Cortex	Neuron	-0.26	0.00087
	Astrocyte	0.24	0.0023
	Oligodendrocyte	0.073	0.36
	Microglia	0.11	0.18
Frontal cortex	Neuron	-0.19	0.031
	Astrocyte	0.2	0.024
	Oligodendrocyte	0.18	0.046
	Microglia	0.082	0.36
Substantia nigra	Neuron	-0.23	0.032
	Astrocyte	0.17	0.1
	Oligodendrocyte	0.11	0.32
	Microglia	0.091	0.4
Anterior cingulate cortex	Neuron	-0.22	0.016
	Astrocyte	0.2	0.031
	Oligodendrocyte	0.16	0.076
	Microglia	0.25	0.0064
Hippocampus	Neuron	-0.33	0.00016
	Astrocyte	0.28	0.0014
	Oligodendrocyte	0.23	0.011
	Microglia	0.43	0.00000066
Cerebellar hemisphere	Neuron	0.045	0.6
	Astrocyte	0.0013	0.99
	Oligodendrocyte	-0.31	0.00025
	Microglia	0.03	0.73
Spinal cord	Neuron	-0.29	0.0053
	Astrocyte	0.16	0.12
	Oligodendrocyte	0.097	0.36
	Microglia	0.3	0.004

7.2.2. Cell type deconvolution of the brain in Alzheimer's disease

To ascertain whether there was an alteration in the relative proportions of the major cell types in the brain of AD patients, we performed cell type deconvolution in a bulk RNA-sequencing dataset of *post-mortem* fusiform gyrus of AD patients and non-AD samples. According to the previous analysis, there is a tendency for a decreased neuron proportion, concomitant with an increase in glia cell proportion, with ageing, in the cerebral cortex. In accordance with previous findings (Mukhin, Pavlov and Klimenko 2017), this shift is accentuated in AD brain relative to non-AD brains. The latter were, in general, older than AD patients, allowing us to rule out age as a possible confounding factor (Figure 7.19).

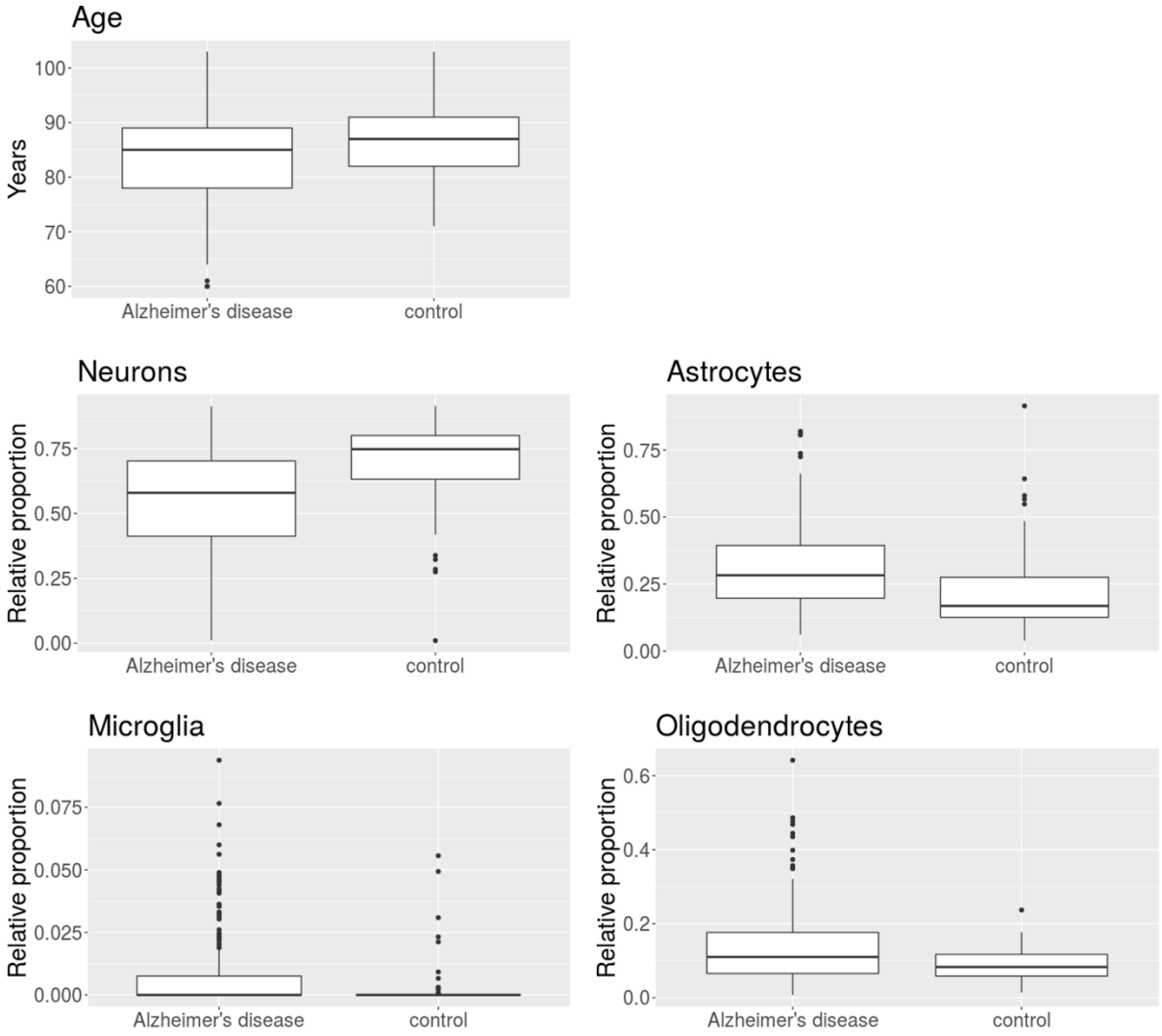

Figure 7.19:Box plots of age and relative proportions of brain cell types, in Alzheimer's disease affected fusiform gyrus and healthy fusiform gyrus (control).

7.2.3. Cell type deconvolution of the brain in Parkinson's disease

PD has been previously associated with a decrease in the number of neurons in the brain, particularly dopaminergic neurons in the *substantia nigra pars compacta*. However, symptoms such as tremors and dementia indicate the occurrence of progressive neurodegeneration in the cerebral cortex in PD brain, which is accentuated in later stages of the disease (Cechetto and Jog 2017; Yau et al. 2018).

The following analysis consisted on performing cell type deconvolution in a bulk RNA-sequencing dataset of *post-mortem* frontal cortex brain tissue, of PD and non-PD samples. Although the shift in the proportion of neurons in this dataset is not as pronounced as in the AD analysis, there is also a decrease in the proportion of neurons in PD frontal cortex relative to non-PD samples, concomitant with an increase in the proportion of astrocytes (Figure 7.20).

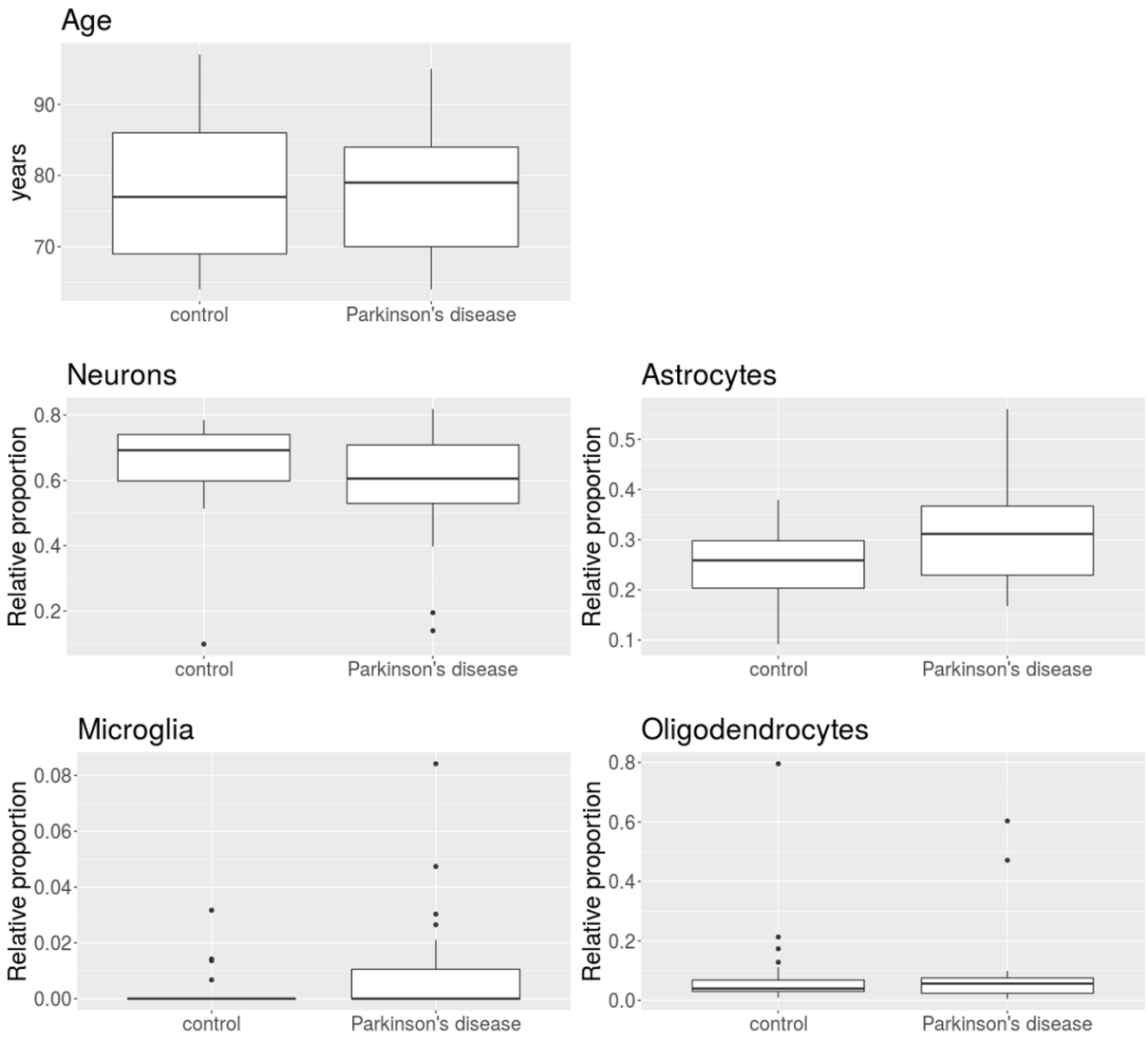

Figure 7.20: Box plots of age and relative proportions of brain cell types, in Parkinson's disease affected frontal cortex and healthy frontal cortex (control).

7.2.4. Immune cell type deconvolution of the healthy brain

For many years the CNS has been considered to be immune privileged, having physical barriers that prevent the entry of blood components, such as leucocytes (Loeveau, Harris and Kipnis 2015). However, it is known that immune cells have important roles in the development, maintenance and repair of the brain. There have been studies showing that although the healthy brain is constituted by about 10% microglia, there is still a need for macrophages to enter and aid in the case of a brain lesion (Tanabe and Yamashita 2018; Shechter et al. 2013).

CD4+ T cells have been described as being key regulators of this monocyte-derived macrophage entry in the CNS, by remotely producing pro-inflammatory cytokines that activate the production of cell-adhesion molecules in the blood-cerebrospinal-fluid (CSF) barrier (Kunis et al. 2013). However, with ageing, there is an increase in the permeability of the BBB, which allows the pathological entry of leucocytes in the CNS (Yamazaki and Kanekiyo 2017). Thus, there is a reduction in the number of T cells, which leads to under-activation of the CSF-barrier, preventing macrophages from entering the brain and removing toxic accumulating plaques (Kunis et al. 2013).

In order to investigate whether the brain is indeed immune privileged, or if there are infiltrating immune cells in the brain, we performed immune cell type deconvolution of bulk RNA-seq datasets of healthy brain tissue, available from GTEx (The GTEx Consortium, 2013), using the whole set of brain tissues available. Given that immune cells do not comprise the major part of cell types in the brain, the analysis focused on studying the absolute levels of immune cells, rather than the relative proportion, using the previously validated immune cell type signature LM22.

The results suggested that the brain is not depleted from immune cells, besides microglia (Figure 7.21). Resting CD4+ memory T cells presented the highest absolute scores, with a relatively high level of variability between individuals. Although T cells have been previously described in the human brain parenchyma, they were found to be mostly memory CD8+ T cells, while memory CD4+ T cells are rarer (Smolders et al. 2018). However, an increase in the transmigration of CD4+ T cells in the CNS has been previously reported in brains affected by AD and PD, comparatively to brains from controls without neurological disease (Gemechu and Bentivoglio 2012). All of the above considered, there is still a need to pursue a systematic analysis of T cell infiltration in the non-diseased human brain at different ages, versus brains affected by neurodegenerative disease (Gemechu and Bentivoglio 2012).

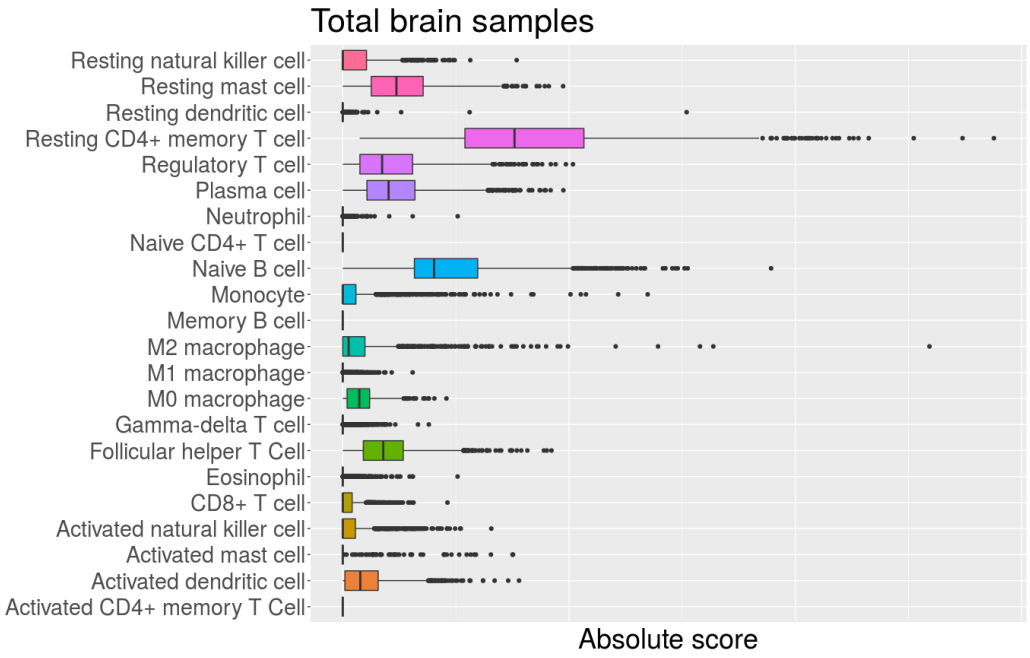

Figure 7.21: Box plots of relative proportions of immune cell types in the non-diseased brain.

7.3. Evaluating the cellular diversity and molecular signature of TAMs

7.3.1. Normalization and clustering

In the following analysis, we aimed at exploring breast tumour-infiltrating macrophage diversity, using the scRNA-seq Azizi dataset.

Prior to normalization, we plotted the percentage of variance of the expression values that was explained by library size (*total_counts*), number of unique features detected (*total_features_by_counts*), patient, replicate and original cell type labelling (*azizi_cell_type*). The percentages explained by library size and the number of unique features detected were relatively high, approaching 10% (Figure 7.22A), indicating that there were technical confounding factors. Plotting the same variables after normalization, the percentage of variance explained by these variables was significantly reduced (Figure 7.22B), demonstrating the successful removal of biases originated during the library preparation and sequencing process.

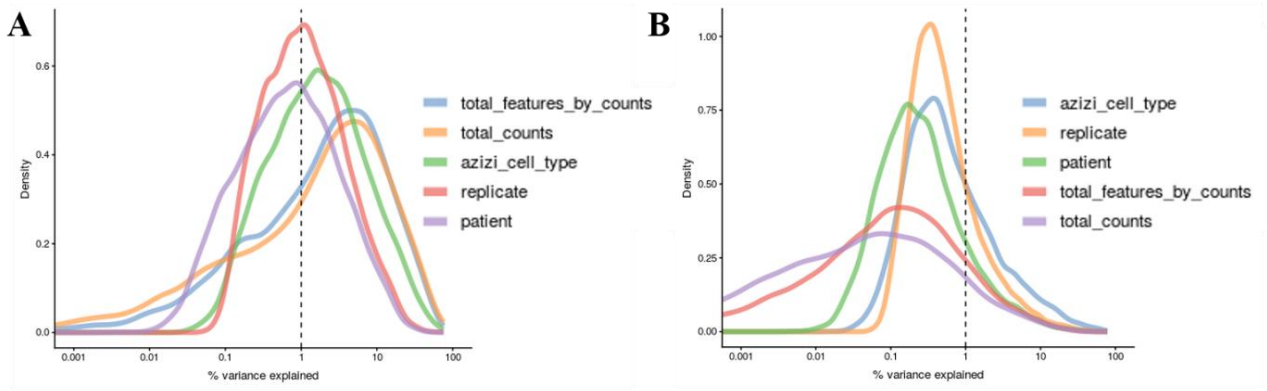

Figure 7.22: Density plots of the percentage of variance explained of the log-expression values across cells. Each curve represents one variable before normalization (A) and after normalization (B).

Plotting the t-SNE coloured by original cell type labelling (Figure 7.23A), and patient (Figure 7.23B), we confirm that cells primarily cluster by cell type. Although there were cell types for which the majority of cells came from only one or two patients, such as natural killer T cells and macrophages, cells that originated from several different patients, e.g., B cells and mast cells, were clustering by cell type. Hence, we decided not to perform patient-batch effect correction.

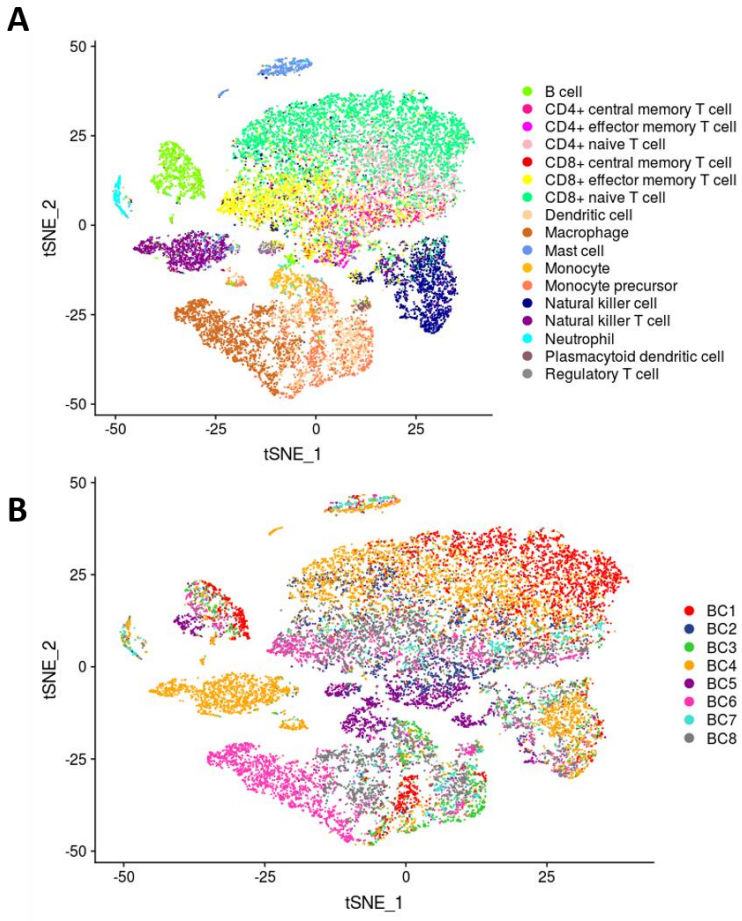

Figure 7.23: t-SNE of the Azizi dataset, coloured by cell type (A) and patient (B).

Next, we performed clustering analysis, to check if the clusters obtained were concordant with cell type. Plotting the t-SNE coloured by the resulting clusters (Figure 7.24B), there were clear groups composed of mainly one cell type, e.g. B cells, mast cells and neutrophils. In general, there was a relatively high concordance between the t-SNE clusters, clusters obtained from the clustering tool, and cell types. The level of concordance between cell types and clusters defined by the Seurat function was summarized in a confusion matrix (Figure 7.24A). For some cell types, there were very homogeneous clusters where the majority of cells belonged to the same type, e.g. neutrophils and mast cells. However, monocytes, together with monocyte precursors, and mDCs, were not easily separated by cell type based solely on gene expression, which is likely depicting the lineage relationship between these cell types, together with macrophages. Hence, there is a likely continuum between these cells rather than discrete groups.

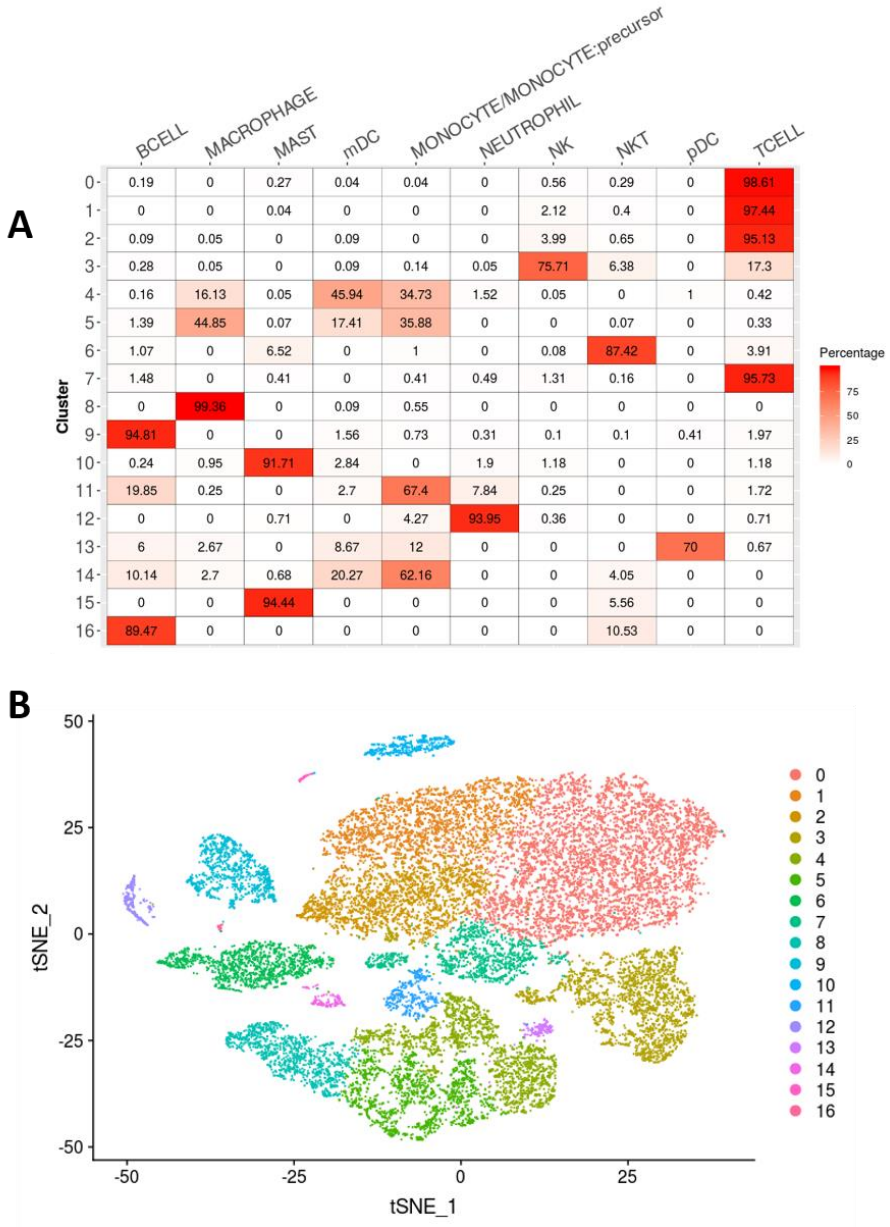

*Figure 7.24: Results from the clustering analysis of breast tumour-infiltrating immune cells. **A** – Confusion matrix showing the percentage of each cell type per cluster. **B** – t-SNE coloured by cluster.*

7.3.2. Identifying macrophage subpopulations

In order to explore the diversity of TAMs, we isolated the macrophages and performed clustering analysis on this subset. We obtained 3 clusters, two from the same patient (BC6) and one from a different patient (BC8) (Figures 7.25A and 7.5B).

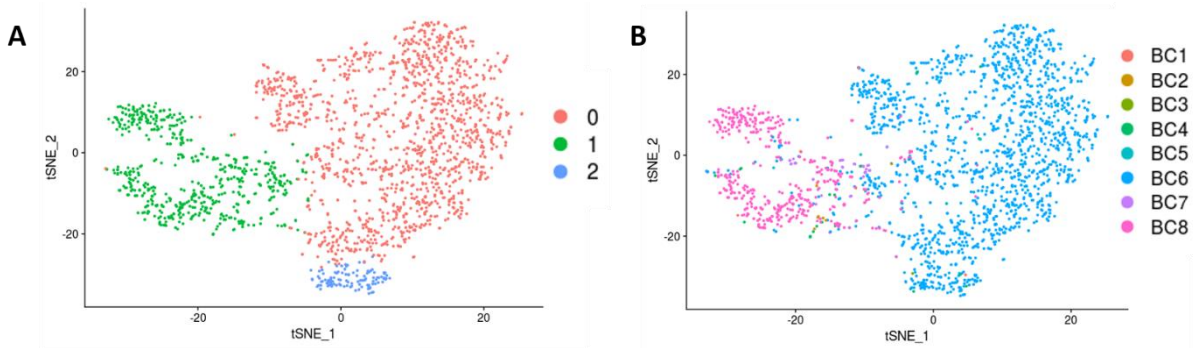

Figure 7.25: t-SNEs of the Azizi dataset, coloured by the results of clustering (A) and by patient (B).

Next, we performed differential expression analysis to identify different macrophage polarizations based on the clusters previously obtained. By comparing each of the clusters to the remaining, we obtained 200 upregulated genes in cluster 0, 289 upregulated genes in cluster 1 and 179 upregulated genes in cluster 2 (average logFC > 0.25 and corrected p-value < 0.05).

Considering the top 10 differentially expressed genes of each cluster (ordered by increasing corrected p-value) (Figure 7.26A), cluster 0 presented genes related with intracellular storage of iron, namely *FTL* (Ferritin Light Chain) and *FTH1* (Ferritin Heavy Chain 1). It is known that macrophages have a pivotal role in iron homeostasis, and that macrophage polarization towards iron sequestration fosters inflammation, in contrast with iron release, which contributes to tissue repair and cell proliferation (Mertens et al. 2016). *FTL* has been previously reported to be upregulated in M1 macrophages, comparatively to M2 macrophages (Mertens et al. 2016).

There were also genes involved in the phagolysosome system, namely *CSTB* (Cystatin B) and *CTSD* (Cathepsin D). *CSTB* codes for a protein that protects against cathepsins that leak out of lysosomes. CSTD is a cathepsin, i.e., a protease that degrades proteins and activates precursors of other bioactive proteins in the lysosome. This enzyme is highly abundant in macrophage lysosomes. The upregulation of this gene could be indicative of a high level of phagocytosis (Bewley et al. 2011), which is associated with M1 phenotype (Atri, Guerfali and Laouini 2018; Weagel et al., 2015). Furthermore, CSTB has been previously reported to be expressed in mouse M1 macrophages, under specific stimulating conditions (Torre-Minguela et al. 2016). Interestingly, cystatins have also been shown to induce production of nitric oxide (NO) in mouse macrophages (Luciano-Montalvo and Meléndez, 2009; Verdot et al. 1996), which, together with the production of ROS, is a hallmark of the M1-phenotype (Atri, Guerfali and Laouini 2018; Weagel et al., 2015). Another upregulated gene in this cluster was *ACP5* (acid phosphatase 5, tartrate resistant), an enzyme that is highly expressed by activated macrophages. Although its function remains unclear, it has been proposed to be involved in the generation of ROS and increased bacterial killing in macrophages (Räisänen et al. 2005).

On the other hand, cluster 2 presented upregulated genes that have been described in the M2-phenotype, e.g., *SEPP1* (Chinetti-Gbaguidi and Staels 2011). Although not present in the top 10, two known M2 macrophage canonical markers were also significantly upregulated in this cluster, *MRC1* (CD206) (Martinez and Gordon 2014) and *CD209* (Lugo-Villarino et al. 2018). While cluster 0 presented genes associated with iron sequestration, macrophages from cluster 2 presented upregulated expression of *SLC40A1*, a gene that codes for the Ferroportin protein, the only known iron exporter. This iron-release macrophage M2-like phenotype has been previously described (Mertens et al. 2016), and is an active topic of research, considering that the presence of iron in the microenvironment stimulates tumour growth and progression. Thus, applying macrophage-targeted chelation strategies might prevent tumour progression, posing new possibilities in the treatment of cancer (Mertens et al. 2016). Another interesting upregulated gene is *LYVE1*, which codes for a cell surface receptor usually expressed on lymphatic endothelial cells. There have been studies linking this gene to the role of M2 macrophages in lymphangiogenesis (Corliss et al. 2016), although in the context of cancer, one study reported an association between *LYVE1*-expressing M2 macrophages and inhibition of melanoma cell proliferation (Dollt et al. 2017). Cluster 2 also expressed genes related with the classical complement pathway, *C1QC* and *C1QA*. In the context of macrophage polarization, C1Q has been described as being overexpressed in M2 macrophages (Fraser et al. 2015).

Finally, cluster 1 analysis did not result in marker genes characteristic of the traditional M1/M2 macrophage classification, being particularly enriched in genes that code for MHC (major histocompatibility complex) class II and ribosomal proteins. To see if these macrophages were more similar to cluster 0 (M1-like) or cluster 2 (M2-like), we performed differential expression analysis comparing only cluster 0 and 2, and plotted the heatmap of the top 10 differentially expressed genes, in order to verify if cluster 1 presented a gene expression profile more similar to either one of these clusters (Figure 7.26B). The expression profile of cluster 1 did not seem to resemble neither cluster 0 or cluster 2, ascertaining it as a different phenotype, possibly more transcriptionally active than the remaining.

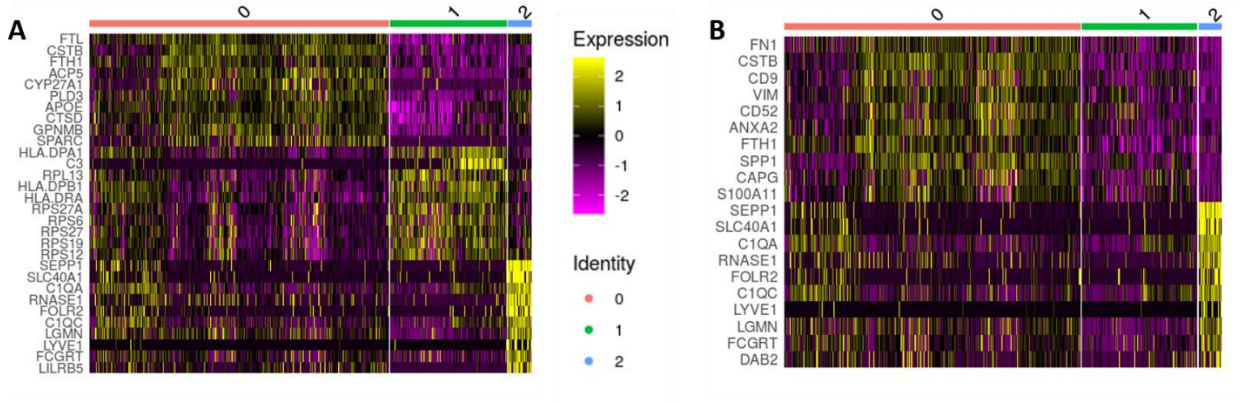

*Figure 7.26: Heatmaps representing the top 10 differentially expressed genes of each cluster from the differential expression analysis between cluster 0, 1 and 2 (**A**), and between cluster 0 and 2 (**B**).*

In summary, this analysis, together with growing information indicating that recognition receptors, cytokines, and the signalling and genetic programs are all factors that influence macrophage polarization, emphasizes the need to recognize a broader functional repertoire for macrophages (Martinez and Gordon 2014). There is also a necessity to carefully determine the effect of particular

TAM subpopulations on tumour progression, before establishing a treatment to target these immune cells (Dollt et al. 2017).

7.4. Understanding how the intra-tumoural diversity and functionality of infiltrating immune cell types is associated with age and prognosis

7.4.1. Deconvolution of immune cell types of TCGA samples

Using the previously validated immune gene signature LM22 (Chen et al. 2018), we performed immune cell type deconvolution of over 800 bulk RNA-seq samples of breast tumour.

By looking at the averages of the relative abundances of each major immune cell type, we confirm that these are in accordance with what is expected from the literature (Cassetta and Pollard 2018), namely macrophages representing the major tumour-infiltrating population, followed by T and B cells (Figures 7.27A and 7.27B). When grouping samples by patient age range, we observe an increase in pro-tumourigenic (M2) macrophages with age, and also a subtle decrease of infiltrating CD8+ T cells (Figures 7.27C and 7.27D). The correlation between age and the increase in the relative abundance of tumour-infiltrating M2 macrophages was significant (Spearman's correlation coefficient, p-value < 0.0001), albeit the correlation coefficient being low (rho = 0.14). Regarding the association between the relative abundance of CD8+ T cells with age, the correlation was not significant at a level of 0.05 (p-value = 0.07 and rho = -0.066) (Figure 7.28). Nevertheless, considering the dimension of the dataset, and the fact that we would not expect dramatic shifts in the proportion of immune cells with ageing, the observed differences could still be worth exploring in future work.

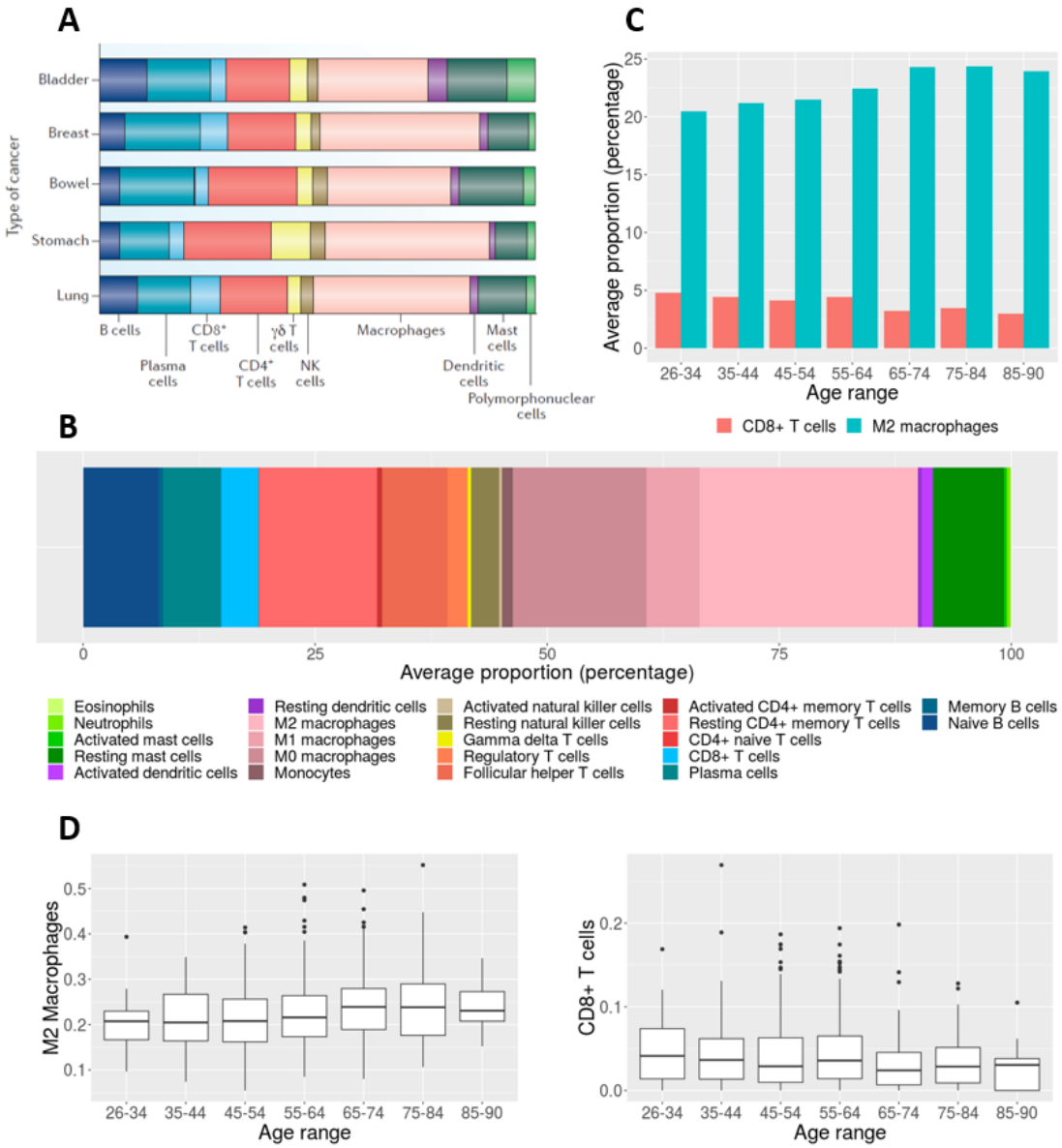

*Figure 7.27: Analysis of cellular composition deconvolution of TCGA breast cancer RNA-seq samples. **A** - Average immune cell composition obtained from the literature, expressed as estimated fractions of leukocyte RNA, in bladder, breast, bowel, stomach and lung cancers. Adapted from Cassetta and Pollard 2018. **B** – Average proportion of immune cell types in breast cancer, obtained from our preliminary analysis. **C** – Average proportion of pro-tumour (M2) macrophages and CD8+ T cells in breast cancer, discriminated by age range. **D** – Box-plots showing the distribution of the fraction of pro-tumour macrophages and CD8+ T cells discriminated by age range.*

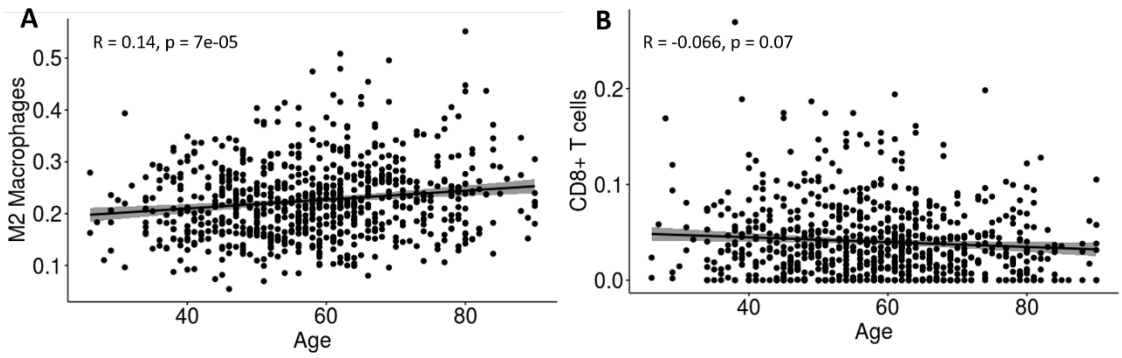

Figure 7.28: Scatter plots of the relative proportion of M2 macrophages vs age (A) and the relative proportion of CD8+ T cells vs age (B). Correlation was assessed using Spearman's Rank-Order Correlation test.

We also found that the relative abundance of CD8+ T cells negatively correlates with the relative abundance of M2 macrophages (rho = -0.3, p-value < 2.2e-16). There is also a positive correlation between the relative abundance of CD8+ T cells and M1 macrophages (rho = 0.39, p-value < 2.2e-16), and a negative correlation between the relative abundance of M1 and M2 macrophages (rho = -0.43, p-value < 2.2e-16). However, given that macrophages comprised the major part of breast tumour-infiltrating immune cells, the decrease in the relative abundance of CD8+ T cells might be biased, i.e., when one population increases, there must be a relative decrease in other immune cell populations.

We performed survival analysis to check if the tumour burden was associated with CD8+ T cell and M2 macrophage relative abundances, finding that a high proportion of M2 macrophages was significantly associated with worse prognosis ($p < 0.0001$, log-rank test), while a high proportion of CD8+ T cells was significantly associated with better prognosis ($p = 0.002$, log-rank test) (Figure 7.29).

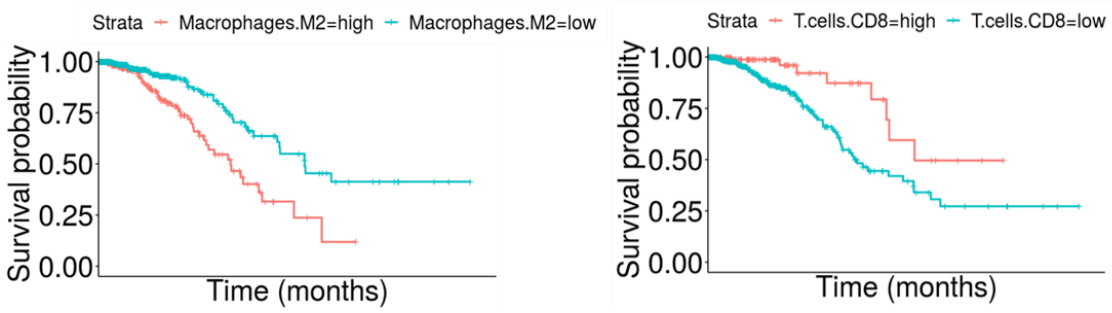

Figure 7.29: Kaplan-Meier plots for patient stratification based on the relative proportion of M2 macrophages (left) and the relative proportion of CD8+ T cells (right).

7.4.2. Differential expression and Gene Set Enrichment analysis

In order to identify cellular pathways strongly associated with differences in the relative proportion of CD8+ T cells and M2 macrophages, possibly related with phenotypical changes in the breast tumour mass, we performed differential expression between two groups of tumours: those with relatively high proportion of CD8+ T cells and low proportion of M2 macrophages, and those with relatively high proportion of M2 macrophages and low proportion of CD8+ T cells. This resulted in 434 upregulated and 450 downregulated genes in the group of samples with relatively high proportion of CD8+T cells

and low proportion of M2 macrophages, after excluding marker genes belonging to the gene signature applied in the deconvolution analysis.

We highlighted in red 3 genes of each category in the volcano plot (*HLA.DOB*, *BCL11A* and *C4orf7* in the upregulated group of genes, and *CLEC5A*, *MMP3* and *METTL11B* in the downregulated group), which presented a high B-statistic value together with a high log fold-change (Figure 7.30).

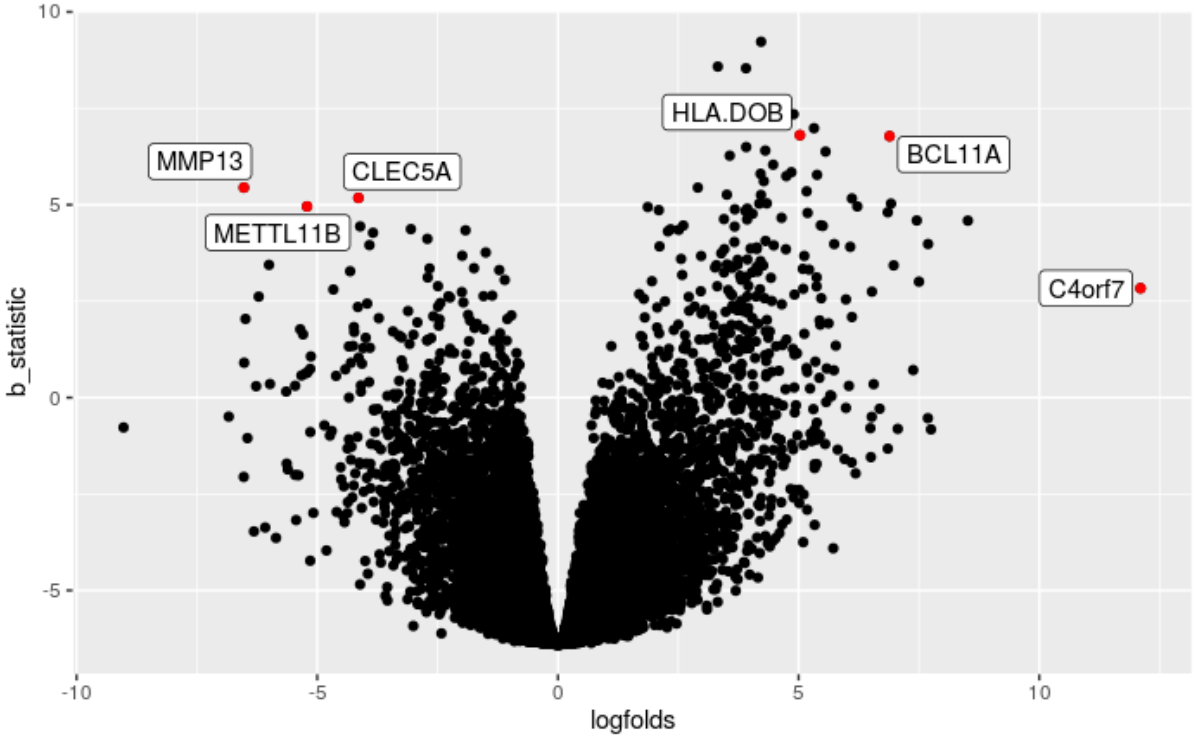

Figure 7.30: Volcano plot showing the results of linear regression analysis comparing the groups of samples with relatively high proportion of CD8+ T cells and low proportion of M2 macrophages vs samples with relatively low proportion of CD8+ T cells and M2 macrophages. Genes were highlighted based on log fold-change values (logfolds) and B-statistic (b_statistic), except for C4orf7, considered an interesting outlier based on a high log fold-change value.

C4orf7, renamed as follicular dendritic cell secreted protein (*FDCSP*), is expressed by follicular dendritic cells and activated leukocytes in the tonsils during immune responses. Interestingly, it has been described as having weak or no expression in peripheral blood leukocytes, spleen and bone marrow (OMIM Entry - * 607241 2019). The FCCSP secreted peptide is known to bind to B cells, and this process is enhanced by T cell-dependent activation signals, albeit the molecular basis for follicular dendritic cell and B cell interactions remains poorly characterized (Marshall et al. 2002). *FDCSP* has been found to have a high level of expression in some cancers, including breast carcinoma, and is thought to contribute to tumour metastases by promoting cancer cell migration and invasion. However, in the context of cancer, it appears to be expressed not by immune, but possibly cancer cells (Wang et al. 2010).

BCL11A codes for a transcription factor highly expressed in the brain, B-lymphocytes, pDCs and the adult erythroid lineage (OMIM Entry - * 606557 2019; Ippolito et al. 2014). The BCL11A transcription factor is required for B cell and pDC development (Lee et al. 2017). It also appears to be related with NK and T cell development, but not with the myeloid compartment (Yu et al. 2012). Regarding cancer, *BCL11A* has been reported to be overexpressed in TNBC and have a role in the expression of expression

of extracellular matrix genes and in promoting tumour development and metastatic progression (Khaled et al. 2015; Seachrist 2018).

HLA-DOB codes for a protein belonging to the class II major histocompatibility complex (MHC) molecules, and is expressed in APCs, namely, B cells, DC and macrophages (Ncbi.nlm.nih.gov 2019). IFN-γ signalling may also induce expression of MHC-II in tumour cells, and it has been reported that high MHC-II expression in TNBC is associated with larger amounts of TILs, and with better disease-free survival in patients who had lymph node metastasis (Park et al. 2017).

MMP13 codes for an enzyme produced predominantly by connective tissue cells. It is involved in the degradation of the major components of the extracellular matrix (OMIM Entry - * 600108 2019). Matrix metalloproteinases are major factors involved in the development of the tumour microenvironment, cancer progression and metastasis. *MMP13* in particular has been described as being highly overexpressed in breast cancer tissue, with a potential role in breast cancer metastasis (Chang et at. 2009; Kotepui et al. 2016).

CLEC5A codes for a cell surface receptor that is strongly involved in the activation and differentiation of myeloid cells, namely macrophages and neutrophils. Being associated with mature stages of myeloid differentiation, *CLEC5A* is expressed constitutively at very low levels, and is highly increased in activated macrophages during infection (Batliner et al. 2011).

METTL11B codes for an enzyme that catalyses the methylation of proteins, e.g. histone proteins of chromatin, regulating gene transcription (Copeland 2018). These enzymes may also be involved in the methylation of non-histone proteins, with a role in cancer development and progression (Copeland 2018; Hamamoto and Nakamura 2016).

In summary, the preliminary investigation of a small subset of significantly differentially expressed genes showed an association of adaptive immune cell activation with the high CD8+ T cell/ low M2 macrophage group, and activation of myeloid cells in the low CD8+ T cell/ high M2 macrophage group.

Next, we focused our analysis on the changes of expression in groups of biologically related genes by performing GSEA. The significant results (FDR < 0.1) of this method are summarized in table 7.3.

Table 7.3: MSigDB's Hallmark Gene Sets (FDR < 0.1).

Gene set	Brief description	FDR
High CD8+ T cells and low M2 macrophages		
HALLMARK_ALLOGRAFT_REJECTION	Genes up-regulated during transplant rejection.	0.000
HALLMARK_INTERFERON_GAMMA_RESPONSE	Genes up-regulated in response to IFNγ.	0.000
HALLMARK_INTERFERON_ALPHA_RESPONSE	Genes up-regulated in response to alpha interferon proteins.	0.005
HALLMARK_TNFA_SIGNALING_VIA_NFKB	Genes regulated by NF-kB in response to TNF	0.082
Low CD8+ T cells and high M2 macrophages		
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	Genes defining epithelial-mesenchymal transition, as in wound healing, fibrosis and metastasis.	0.000
HALLMARK_COAGULATION	Genes encoding components of blood coagulation system; also up-regulated in platelets.	0.058
HALLMARK_UV_RESPONSE_DN	Genes down-regulated in response to ultraviolet (UV) radiation.	0.058
HALLMARK_TGF_BETA_SIGNALING	Genes up-regulated in response to TGF-β1.	0.058
HALLMARK_ESTROGEN_RESPONSE_EARLY	Genes defining early response to oestrogen.	0.076
HALLMARK_ESTROGEN_RESPONSE_LATE	Genes defining late response to oestrogen.	0.091

One of the most significant hallmarks obtained for the high CD8+ T cell/ low M2 macrophage group was allograft rejection. Allograft rejection consists of an elicited response of the adaptive immune system to a donor tissue, in which T cells react to allogeneic MHC molecules expressed on the surface of donor cells. The emergence of this hallmark may be related with the immune system recognizing tumour antigens as nonself, which would elicit a cytotoxic T-cell attack on tumour cells (Genome.jp. 2019). The remaining hallmarks were related with the production of pro-inflammatory cytokines. IFN-γ, in particular, is secreted by NK and T cells upon activation of antigen-specific immunity. In the tumour microenvironment, this cytokine is associated with anti-proliferative, pro-apoptotic and anti-tumour mechanisms (Boehm et al. 1997).

Regarding the low CD8+ T cell/ high M2 macrophage group, the most significant hallmark was epithelial-mesenchymal transition, a process through which cells lose adhesion and gain migratory capacities. This mechanism is involved in tumour progression with metastatic expansion (Roche 2018). Another significant hallmark was TGF-β signalling, a cytokine previously implicated in the polarization of macrophages towards the M2-phenotype (Zhang et al. 2016). TGF-β also converts effector T-cells, which normally attack tumour cells, into regulatory (suppressor) T-cells, which turn off the inflammatory reaction (Dahmani and Delisle 2018). These results are concordant with the hypothesis that macrophage polarization is determined by microenvironment stimuli, M2 macrophages can affect malignancy progression and metastasis, and are involved in immune suppression (Weagel et al., 2015).

8. Concluding remarks

Human bodies are frequently said to have 210 different types of cells (Trapnell, 2015). However, a single cell type can be divided by functional differences and subcategorized by unique gene expression programs (Trapnell, 2015). Advances in scRNA-seq technologies have revealed different gene expression profiles between cells once categorized as the same type, calling into question how we define cell type in the first place (Gage, Linker and Bedrosian, 2019).

Wagner's group defined cell identity as the outcome of the instantaneous intersection of all factors that affect it (Wagner, Regev and Yosef, 2016). Type is the permanent aspect of a cell, and states arise transiently during time-dependent processes, such as the cell cycle (Wagner, Regev and Yosef, 2016). Therefore, we can think of marker genes as the set of genes that are similarly expressed across all cells with identical function and that are consistent across all states (Gage, Linker and Bedrosian, 2019). Traditional approaches to cell type identification were based on morphological characterization, and/or identifying the presence of a small set of known markers. scRNA-seq methods, on the other hand, enable cell type classification with little or no prior knowledge, possibly revealing previously unidentified variations in cellular phenotypes across numerous tissue types (Gage, Linker and Bedrosian, 2019). In this way, scRNA-seq provides a conceptual framework by which to assess cell type (Gage, Linker and Bedrosian, 2019).

Rapid progress in the development of scRNA-seq protocols and computational methods in recent years has provided many valuable insights into the diversity of the immune system, in health and disease (Stubbington et al. 2017).

In this work, we focused on implementing scRNA-seq data analysis approaches to unravel the role of the immune response in ageing-related human diseases, namely neurodegenerative and oncological ones.

The characterization of gene expression signatures of the major cell types of the brain enabled the estimation of relative proportions of neurons and glia cells in brain tissue. This allowed us to look at the differences in these proportions across different CNS areas, as well as assess the correlation between neuronal loss and glial proliferation in brain tissue and ageing, and the accentuation of this phenomena in AD and PD.

The use of scRNA-seq has also transformed cancer research by enabling the estimation of immune cell content in tumour tissues and the profiling of gene expression for tumour-infiltrating immune cells. In this work, focusing on breast cancer, we performed an inspection of the tumour microenvironment and its consequences for prognosis. We estimated the relative abundances of tumour-infiltrating immune cell types and how they correlated with age. Furthermore, we ascertained how the variability of tumour-infiltrating macrophage and T cell patterns is associated with either induced immune suppression, or increased immunosurveillance in the tumour microenvironment.

Finally, the immune system comprises a variety of immune cells, not only of multiple cell types, but also of different states within a cell type. Characterizing this complexity requires studies at the single-cell resolution. Focusing on TAMs, due to their variety of microenvironment stimuli-dependent polarizations, we were able to characterize different activation states in the breast TME, which encompassed classically activated macrophages (M1), alternatively activated macrophages (M2) and a transcriptionally activated subtype of macrophage, not previously identified to the extent of our knowledge.

8.1. Analysis limitations

Single-cell transcriptomics is a relatively new and still growing field. Hence, a gold standard for a scRNA-seq data analysis pipeline is yet to come. Taking this into consideration, during this work we tested different combinations of methods to perform feature selection and obtain cell type marker genes. Although some methods exhibited significant overlapping results, different methods usually resulted in different gene signatures, which led to different cell type deconvolution scores for the same dataset. This emphasizes the need to extensively benchmark scRNA-seq analysis pipelines.

There are also different cell type deconvolution algorithms, each with different strengths and limitations (Li et al. 2017; Qiao et al. 2012; Racle et al. 2017; Finotello et al. 2019). In this work, we selected CIBERSORT for its extensive validation on deconvolution of immune subsets, particularly in the context of tumour content (Finotello and Trajanoski 2018). However, the accuracy of this tool may be affected by not taking into account cross-subject heterogeneity in cell type-specific gene expression as well as within-cell type stochasticity of single-cell gene expression (Wang et al. 2019). In addition, the expression of single cells can be different between their physiological context (i.e., in a tissue) and when they are extracted for sequencing.

The results obtained and validated in independent datasets clearly state that there is a need to increase the biological variability of individuals in single-cell data. Moreover, all of the derived signatures were from a specific brain region. Given the potential of scRNA-seq data analyses to go deeper in the knowledge of specific cell types and states, cell type deconvolution analyses should be performed using the same tissue, in order to discard potential tissue-site bias in the signature.

Another caveat in our analysis was related with the limited number of samples of patients from young age-ranges, particularly in the GTEx dataset, given that we were working with *post-mortem* healthy brain tissue. To validate our results, in the future, we will need to integrate different datasets in order to obtain a more comprehensive set of brain tissue samples, equally distributed from young adulthood to old age.

8.2. Future perspectives

The remarkable heterogeneity of tumour-infiltrating myeloid cells may conceal the real extent of anti-tumour effectors and the molecular determinants that control their functions (Kiss et al. 2018). Thus, in order to further consolidate our work, we propose to overcome this limitation by bringing individual myeloid cell type and state characterization enabled by comparing single-cell transcriptomes.

Karine Serre's team at iMM has established a mouse model of tumour regression strictly dependent on the presence of myeloid cells, especially macrophages. This is based on the myeloid cell properties to respond to stimulatory agents. To further evaluate the cellular diversity and molecular signature of myeloid cells with anti-tumour functions, we will generate a scRNA-seq dataset consisting of a mouse syngeneic TNBC cell line model injected in the mammary fat pad of mice. We will sort samples of tumours with infiltrating anti-tumour myeloid cells (inducing tumour regression) and pro-tumour myeloid cells (promoting tumour growth), with particular interest to experimentally-induced macrophages with anti-tumour functions. This will be performed using FACS with a myeloid cell bulk

marker coupled with the 10X Genomics microdroplet technology. After sorting, single cells will be sequenced using an Illumina platform in order to obtain their gene expression profiles.

This dataset will enable us to further characterise subtypes of infiltrating myeloid cells and define their molecular distinctiveness, contributing to the creation of a single-cell expression atlas of myeloid lineage-specific anti-tumour molecular effectors. One of the strengths of our project is that we built it at the interface between human tumour immunity, evaluated through publicly available transcriptome datasets, and the analysis of mouse myeloid cells performing anti-tumour function in breast cancer. This way we can determine regulators (effector molecules or transcription factors) of anti-tumour functions of myeloid cells from the mouse models and assess their prognostic potential in the human data sets. Finally, the characterization of tumour-infiltrating immune cells may disclose better strategies for overcoming immune suppression and restoring immunosurveillance in cancer.

9. References

A

Adamo, A. (2013). Nutritional factors and aging in demyelinating diseases. Genes & Nutrition, 9(1), p.360. doi: 10.1007/s12263-013-0360-8.

Alessandrì, L., Arigoni, M. and Calogero, R. (2019). Differential Expression Analysis in Single-Cell Transcriptomics. Methods in Molecular Biology, pp.425-432. doi: 10.1007/978-1-4939-9240-9_25.

Alkabban, F. and Ferguson, T. (2019). Cancer, Breast. [online] Ncbi.nlm.nih.gov. Available at: https://www.ncbi.nlm.nih.gov/books/NBK482286/ [Accessed 5 Sep. 2019].

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Andrews, T. (2019). M3Drop: Michaelis-Menten Modelling of Dropouts in single-cell RNASeq. R package version 1.10.0. Available online at: https://github.com/tallulandrews/M3Drop.

Andrews, T. and Hemberg, M. (2018). M3Drop: dropout-based feature selection for scRNASeq. Bioinformatics, 35(16), pp.2865-2867. doi.org/10.1093/bioinformatics/bty1044.

Atri, C., Guerfali, F. and Laouini, D. (2018). Role of Human Macrophage Polarization in Inflammation during Infectious Diseases. International Journal of Molecular Sciences, 19(6), p.1801. doi: 10.3390/ijms19061801.

Aysola, K., Desai, A., Welch, C., Xu, J., Qin, Y., Reddy, V., Matthews, R., Owens, C., Okoli, J., Beech, D., Piyathilake, C., Reddy, S., Rao, V. N. (2013). Triple Negative Breast Cancer - An Overview. Hereditary genetics: current research, 2013(Suppl 2), p.001. doi:10.4172/2161-1041.s2-001.

Azizi, E., Carr, A., Plitas, G., Cornish, A., Konopacki, C., Prabhakaran, S., Nainys, J., Wu, K., Kiseliovas, V., Setty, M., Choi, K., Fromme, R., Dao, P., McKenney, P., Wasti, R., Kadaveru, K., Mazutis, L., Rudensky, A. and Pe'er, D. (2018). Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. Cell, 174(5), pp.1293-1308.e36.doi: 10.1016/j.cell.2018.05.060.

B

Baran-Gale, J., Chandra, T. and Kirschner, K. (2017). Experimental design for single-cell RNA sequencing. Briefings in Functional Genomics, 17(4), pp.233-239. doi: 10.1093/bfgp/elx035.

Barry, D., Pakan, J. and McDermott, K. (2014). Radial glial cells: Key organisers in CNS development. The International Journal of Biochemistry & Cell Biology, 46, pp.76-79. doi.org/10.1016/j.biocel.2013.11.013.

Batliner, J., Mancarelli, M., Jenal, M., Reddy, V., Fey, M., Torbett, B. and Tschan, M. (2011). CLEC5A (MDL-1) is a novel PU.1 transcriptional target during myeloid differentiation. Molecular Immunology, 48(4), pp.714-719. doi: 10.1016/j.molimm.2010.10.016. Benz, C. (2008). Impact of aging on the biology of breast cancer. Critical Reviews in Oncology/Hematology, 66(1), pp.65-74.

Benz, C. (2008). Impact of aging on the biology of breast cancer. Critical Reviews in Oncology/Hematology, 66(1), pp.65-74.

Bewley, M., Pham, T., Marriott, H., Noirel, J., Chu, H., Ow, S., Ryazanov, A., Read, R., Whyte, M., Chain, B., Wright, P. and Dockrell, D. (2011). Proteomic Evaluation and Validation of Cathepsin D Regulated Proteins in Macrophages Exposed toStreptococcus pneumoniae. Molecular & Cellular Proteomics, 10(6), pp.M111.008193. doi.org/10.1074/mcp.m111.008193.

Boehm, U., Klamp, T., Groot, M. and Howard, J. (1997). CELLULAR RESPONSES TO INTERFERON-γ. Annual Review of Immunology, 15(1), pp.749-795. doi.org/10.1146/annurev.immunol.15.1.749.

Braak, H. and Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. Acta Neuropathologica, 82(4), pp.239-259. doi.org/10.1007/bf00308809.

Bracci, L., Schiavoni, G., Sistigu, A. and Belardelli, F. (2013). Immune-based mechanisms of cytotoxic chemotherapy: implications for the design of novel and rationale-based combined treatments against cancer. Cell Death & Differentiation, 21(1), pp.15-25. doi: 10.1038/cdd.2013.67.

Bray, N., Pimentel, H., Melsted, P. and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology, 34(5), pp.525-527. doi: 10.1038/nbt.3519.

Brennecke, P., Anders, S., Kim, J., Kołodziejczyk, A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S., Marioni, J. and Heisler, M. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nature Methods, 10(11), pp.1093-1095. doi: 10.1038/nmeth.2645.

C

Cano, L. and Lopera, D. (2013). Introduction to T and B lymphocytes in Anaya J M et al. Autoimmunity: From Bench to Bedside. Bogota (Colombia). El Rosario University Press.

Cassetta, L. and Pollard, J. (2018). Targeting macrophages: therapeutic approaches in cancer. Nature Reviews Drug Discovery, 17(12), pp.887-904. doi: 10.1038/nrd.2018.169.

Cechetto, D. and Jog, M. (2017). Parkinson's Disease and the Cerebral Cortex. The Cerebral Cortex in Neurodegenerative and Neuropsychiatric Disorders, pp.177-193. doi.org/10.1016/B978-0-12-801942-9.00007-0.

Chang, H., Yang, M., Yang, Y., Hou, M., Hsueh, E. and Liu, S. (2009). MMP13 is potentially a new tumor marker for breast cancer diagnosis. Oncology Reports, 22(05), pp.1119-1127.doi.org/10.3892/or_00000544.

Chen, B., Khodadoust, M., Liu, C., Newman, A. and Alizadeh, A. (2018). Profiling Tumor Infiltrating Immune Cells with CIBERSORT. Methods in Molecular Biology, pp.243-259. doi: 10.1007/978-1-4939-7493-1_12.

Cheng, L., Gao, L., Guan, W., Mao, J., Hu, W., Qiu, B., Zhao, J., Yu, Y. and Pei, G. (2015). Direct conversion of astrocytes into neuronal cells by drug cocktail. Cell Research, 25(11), pp.1269-1272. doi.org/10.1038/cr.2015.120.

Chinetti-Gbaguidi, G. and Staels, B. (2011). Macrophage polarization in metabolic disorders. Current Opinion in Lipidology, 22(5), pp.365-372. doi: 10.1097/MOL.0b013e32834a77b4.

Clayton, K., Van Enoo, A. and Ikezu, T. (2017). Alzheimer's Disease: The Role of Microglia in Brain Homeostasis and Proteopathy. Frontiers in Neuroscience, 11, p.680. doi: 10.3389/fnins.2017.00680.

Clough, E. and Barrett, T. (2016). The Gene Expression Omnibus Database. Methods in Molecular Biology, pp.93-110. doi: 10.1007/978-1-4939-3578-9_5.

Copeland, R. (2018). Protein methyltransferase inhibitors as precision cancer therapeutics: a decade of discovery. Philosophical Transactions of the Royal Society B: Biological Sciences, 373(1748), p.20170080. doi.org/10.1098/rstb.2017.0080.

Corliss, B., Azimi, M., Munson, J., Peirce, S. and Murfee, W. (2016). Macrophages: An Inflammatory Link Between Angiogenesis and Lymphangiogenesis. Microcirculation, 23(2), pp.95-121. doi: 10.1111/micc.12259.

Costantini, E., D'Angelo, C. and Reale, M. (2018). The Role of Immunosenescence in Neurodegenerative Diseases. Mediators of Inflammation, 2018, pp.1-12. doi: 10.1155/2018/6039171.

Cotechini, T., Medler, T. and Coussens, L. (2015). Myeloid Cells as Targets for Therapy in Solid Tumors. The Cancer Journal, 21(4), pp.343-350. doi: 10.1097/ppo.0000000000000132.

Crowell, H., Soneson, C., Germain, P., Calini, D., Collin, L., Raposo, C., Malhotra, D. and Robinson, M. (2019). On the discovery of population-specific state transitions from multi-sample multi-condition single-cell RNA sequencing data. doi: https://doi.org/10.1101/713412.

D

Dahmani, A. and Delisle, J. (2018). TGF-β in T Cell Biology: Implications for Cancer Immunotherapy. Cancers, 10(6), p.194. doi: 10.3390/cancers10060194.

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., & Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. American journal of cancer research, 5(10), pp.2929–2943.

Darmanis, S., Sloan, S., Zhang, Y., Enge, M., Caneda, C., Shuer, L., Hayden Gephart, M., Barres, B. and Quake, S. (2015). A survey of human brain transcriptome diversity at the single cell level. Proceedings of the National Academy of Sciences, 112(23), pp.7285-7290. doi.org/10.1073/pnas.1507125112.

Delves, P. and Roitt, I. (2000). The Immune System. New England Journal of Medicine, 343(1), pp.37-49. doi: 10.1056/nejm20000706343010.

Dobin, A., Davis, C., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. (2012). STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), pp.15-21. doi: 10.1093/bioinformatics/bts635.

Dollt, C., Becker, K., Michel, J., Melchers, S., Weis, C., Schledzewski, K., Krewer, A., Kloss, L., Gebhardt, C., Utikal, J. and Schmieder, A. (2017). The shedded ectodomain of Lyve-1 expressed on M2-like tumor-associated macrophages inhibits melanoma cell proliferation. Oncotarget, 8(61). doi: 10.18632/oncotarget.21771.

Dossi, E., Vasile, F. and Rouach, N. (2018). Human astrocytes in the diseased brain. Brain Research Bulletin, 136, pp.139-156. doi: 10.1016/j.brainresbull.2017.02.00.

Dumitriu, A., Golji, J., Labadorf, A., Gao, B., Beach, T., Myers, R., Longo, K. and Latourelle, J. (2015). Integrative analyses of proteomics and RNA transcriptomics implicate mitochondrial processes, protein folding pathways and GWAS loci in Parkinson disease. BMC Medical Genomics, 9(1). doi: 10.1186/s12920-016-0164-y.

EEE

Elliott, L., Doherty, G., Sheahan, K. and Ryan, E. (2017). Human Tumor-Infiltrating Myeloid Cells: Phenotypic and Functional Diversity. Frontiers in Immunology, 8, p.86. doi: 10.3389/fimmu.2017.00086.

Erkkinen, M., Kim, M. and Geschwind, M. (2017). Clinical Neurology and Epidemiology of the Major Neurodegenerative Diseases. Cold Spring Harbor Perspectives in Biology, 10(4), p.a033118. doi: 10.1101/cshperspect.a033118.

F

Farias, G., Cornejo, A., Jimenez, J., Guzman, L. and B. Maccioni, R. (2011). Mechanisms of Tau Self-Aggregation and Neurotoxicity. Current Alzheimer Research, 8(6), pp.608-614. doi: 10.2174/156720511796717258.

Farina, C., Aloisi, F. and Meinl, E. (2007). Astrocytes are active players in cerebral innate immunity. Trends in Immunology, 28(3), pp.138-145. doi.org/10.1016/j.it.2007.01.005.

Feher, J. (2017). Quantitative human physiology. 2nd ed. Elsevier.

Fentiman, I. and D'Arrigo, C. (2004). Pathogenesis of breast carcinoma. International Journal of Clinical Practice, 58(1), pp.35-40. doi.org/10.1111/j.1368-5031.2003.0091.x.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A., Slichter, C., Miller, H., McElrath, M., Prlic, M., Linsley, P. and Gottardo, R. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biology, 16(1). doi: 10.1186/s13059-015-0844-5.

Finotello, F., Mayer, C., Plattner, C., Laschober, G., Rieder, D., Hackl, H., Krogsdam, A., Loncova, Z., Posch, W., Wilflingseder, D., Sopper, S., Ijsselsteijn, M., Brouwer, T., Johnson, D., Xu, Y., Wang, Y., Sanders, M., Estrada, M., Ericsson-Gonzalez, P., Charoentong, P., Balko, J., de Miranda, N. and

Trajanoski, Z. (2019). Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Medicine, 11(1), p.34. doi: 10.1186/s13073-019-0638-6.

Finotello, F. and Trajanoski, Z. (2018). Quantifying tumor-infiltrating immune cells from transcriptomics data. Cancer Immunology, Immunotherapy, 67(7), pp.1031-1040. doi: 10.1007/s00262-018-2150-z.

Franceschi, C., Bonafè, M., Valensin, S., Olivieri, F., De Luca, M., Ottaviani, E. and De Benedictis, G. (2006). Inflamm-aging: An Evolutionary Perspective on Immunosenescence. Annals of the New York Academy of Sciences, 908(1), pp.244-254. doi.org/10.1111/j.1749-6632.2000.tb06651.x.

Fraser, D., Melzer, E., Camacho, A. and Gomez, M. (2015). Macrophage production of innate immune protein C1q is associated with M2 polarization. The Journal of Immunology, 194(1), pp.56.11.

Fridman, W., Galon, J., Dieu-Nosjean, M., Cremer, I., Fisson, S., Damotte, D., Pagès, F., Tartour, E. and Sautès-Fridman, C. (2010). Immune Infiltration in Human Cancer: Prognostic Significance and Disease Control. Current Topics in Microbiology and Immunology, pp.1-24. doi.org/10.1007/82_2010_46.

Fundo IMM Laço - Oncologia e cancro da mama em Portugal. [online] Available at: https://fundoimmlaco.pt/estatisticas/ [Accessed 15 Jul. 2019].

G

Gage, F., Linker, S. and Bedrosian, T. (2019). Opinion: How to Define Cell Type. [online] The Scientist Magazine®. Available at: https://www.the-scientist.com/opinion/opinion-how-to-define-cell-type-30668 [Accessed 26 Sep. 2019].

Gao, H. and Hong, J. (2008). Why neurodegenerative diseases are progressive: uncontrolled inflammation drives disease progression. Trends in Immunology, 29(8), pp.357-365. doi: 10.1016/j.it.2008.05.002.

Gefen, T., Kim, G., Bolbolan, K., Geoly, A., Ohm, D., Oboudiyat, C., Shahidehpour, R., Rademaker, A., Weintraub, S., Bigio, E., Mesulam, M., Rogalski, E. and Geula, C. (2019). Activated Microglia in Cortical White Matter Across Cognitive Aging Trajectories. Frontiers in Aging Neuroscience, 11, p.94. doi.org/10.3389/fnagi.2019.00094.

Gemechu, J. and Bentivoglio, M. (2012). T Cell Recruitment in the Brain during Normal Aging. Frontiers in Cellular Neuroscience, 6. doi.org/10.3389/fncel.2012.00038.

Genome.jp. (2019). KEGG PATHWAY: Allograft rejection - Homo sapiens (human). [online] Available at: https://www.genome.jp/kegg-bin/show_pathway?map=hsa05330&show_description=show [Accessed 21 Sep. 2019].

Goldman, A. and Prabhakar, B. (1996). Immunology Overview. In Baron S (4th ed). Medical Microbiology. Galveston (TX): University of Texas Medical Branch at Galveston.

Gonda, K., Shibata, M., Ohtake, T., Matsumoto, Y., Tachibana, K., Abe, N., Ohto, H., Sakurai, K. and Takenoshita, S. (2017). Myeloid-derived suppressor cells are increased and correlated with type 2 immune responses, malnutrition, inflammation, and poor prognosis in patients with breast cancer. Oncology Letters, 14(2), pp.1766-1774. doi: 10.3892/ol.2017.6305.

Gong, H., Do, D. and Ramakrishnan, R. (2018). Single-Cell mRNA-Seq Using the Fluidigm C1 System and Integrated Fluidics Circuits. Methods in Molecular Biology, pp.193-207. doi: 10.1007/978-1-4939-7834-2_10.

González-Reyes, R., Nava-Mesa, M., Vargas-Sánchez, K., Ariza-Salamanca, D. and Mora-Muñoz, L. (2017). Involvement of Astrocytes in Alzheimer's Disease from a Neuroinflammatory and Oxidative Stress Perspective. Frontiers in Molecular Neuroscience, 10.

H

Hamamoto, R. and Nakamura, Y. (2016). Dysregulation of protein methyltransferases in human cancer: An emerging target class for anticancer therapy. Cancer Science, 107(4), pp.377-384. doi: 10.1111/cas.12884.

Hindle, J. (2010). Ageing, neurodegeneration and Parkinson's disease. Age and Ageing, 39(2), pp.156-161. doi.org/10.1093/ageing/afp223.

Hooke, R. (1665). Micrographia: or Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses. With Observations and Inquiries Thereupon. The Royal Society.

Hol, E. and Pekny, M. (2015). Glial fibrillary acidic protein (GFAP) and the astrocyte intermediate filament system in diseases of the central nervous system. Current Opinion in Cell Biology, 32, pp.121-130. doi.org/10.1016/j.ceb.2015.02.004.

Huang, W., Zhang, X. and Chen, W. (2016). Role of oxidative stress in Alzheimer's disease. Biomedical Reports, 4(5), pp.519-522. doi: 10.3892/br.2016.630.

Huber, W., Carey, J., Gentleman, R., Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods, 12, p.115.

Hume, D., Irvine, K. and Pridans, C. (2019). The Mononuclear Phagocyte System: The Relationship between Monocytes and Macrophages. Trends in Immunology, 40(2), pp.98-112. doi.org/10.1016/j.it.2018.11.007.

Hwang, B., Lee, J. and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Experimental & Molecular Medicine, 50(8), p.96. doi.org/10.1038/s12276-018-0071-8

I

Ippolito, G., Dekker, J., Wang, Y., Lee, B., Shaffer, A., Lin, J., Wall, J., Lee, B., Staudt, L., Liu, Y., Iyer, V. and Tucker, H. (2014). Dendritic cell fate is determined by BCL11A. Proceedings of the National Academy of Sciences, 111(11), pp.E998-E1006. doi.org/10.1073/pnas.1319228111.

J

Joe, E., Choi, D., An, J., Eun, J., Jou, I. and Park, S. (2018). Astrocytes, Microglia, and Parkinson's Disease. Experimental Neurobiology, 27(2), p.77. doi: 10.5607/en.2018.27.2.77.

Jolliffe, I. and Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), p.20150202. doi: 10.1098/rsta.2015.0202.

K

Karlmark, K., Tacke, F. and Dunay, I. (2012). Monocytes in health and disease — Minireview. European Journal of Microbiology and Immunology, 2(2), pp.97-102. doi: 10.1556/eujmi.2.2012.2.1.

Keren-Shaul, H., Spinrad, A., Weiner, A., Matcovitch-Natan, O., Dvir-Szternfeld, R., Ulland, T., David, E., Baruch, K., Lara-Astaiso, D., Toth, B., Itzkovitz, S., Colonna, M., Schwartz, M. and Amit, I. (2017). A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. Cell, 169(7), pp.1276-1290.e17. doi: 10.1016/j.cell.2017.05.018.

Kerkar, S. and Restifo, N. (2012). Cellular Constituents of Immune Escape within the Tumor Microenvironment. Cancer Research, 72(13), pp.3125-3130. doi: 10.1158/0008-5472.can-11-4094.

Khaled, W., Choon Lee, S., Stingl, J., Chen, X., Raza Ali, H., Rueda, O., Hadi, F., Wang, J., Yu, Y., Chin, S., Stratton, M., Futreal, A., Jenkins, N., Aparicio, S., Copeland, N., Watson, C., Caldas, C. and Liu, P. (2015). BCL11A is a triple-negative breast cancer gene with critical functions in stem and progenitor cells. Nature Communications, 6(1). doi.org/10.1038/ncomms6987.

Kiselev, V., Kirschner, K., Schaub, M., Andrews, T., Yiu, A., Chandra, T., Natarajan, K., Reik, W., Barahona, M., Green, A. and Hemberg, M. (2017). SC3: consensus clustering of single-cell RNA-seq data. Nature Methods, 14(5), pp.483-486. doi: 10.1038/nmeth.4236.

Kishore, J., Goel, M. and Khanna, P. (2010). Understanding survival analysis: Kaplan-Meier estimate. International Journal of Ayurveda Research, 1(4), p.274. doi: 10.4103/0974-7788.76794.

Kiss, M., Van Gassen, S., Movahedi, K., Saeys, Y. and Laoui, D. (2018). Myeloid cell heterogeneity in cancer: not a single cell alike. Cellular Immunology, 330, pp.188-201. doi: 10.1016/j.cellimm.2018.02.008.

Kolarova, M., García-Sierra, F., Bartos, A., Ricny, J. and Ripova, D. (2012). Structure and Pathology of Tau Protein in Alzheimer Disease. International Journal of Alzheimer's Disease, 2012, pp.1-13. doi: 10.1155/2012/731526.

Kolodziejczyk, A., Kim, J., Svensson, V., Marioni, J. and Teichmann, S. (2015). The Technology and Biology of Single-Cell RNA Sequencing. Molecular Cell, 58(4), pp.610-620. doi: 10.1016/j.molcel.2015.04.005.

Kotepui, M., Punsawad, C., Chupeerach, C., Songsri, A., Charoenkijkajorn, L. and Petmitr, S. (2016). Differential expression of matrix metalloproteinase-13 in association with invasion of breast cancer. Współczesna Onkologia, 3, pp.225-228. doi: 10.5114/wo.2016.61565.

Kroemer, G., Senovilla, L., Galluzzi, L., André, F. and Zitvogel, L. (2015). Natural and therapy-induced immunosurveillance in breast cancer. Nature Medicine, 21(10), pp.1128-1138. doi: 10.1038/nm.3944.

Kuhn, M. (2019). caret: Classification and Regression Training. R package version 6.0-84. Available online at: https://CRAN.R-project.org/package=caret.

Kunis, G., Baruch, K., Rosenzweig, N., Kertser, A., Miller, O., Berkutzki, T. and Schwartz, M. (2013). IFN-γ-dependent activation of the brain's choroid plexus for CNS immune surveillance and repair. Brain, 136(11), pp.3427-3440. doi: 10.1093/brain/awt259.

Kurosaki, T., Kometani, K. and Ise, W. (2015). Memory B cells. Nature Reviews Immunology, 15(3), pp.149-159. doi.org/10.1038/nri3802.

L

Lai, C., Guo, S., Cheng, L. and Wang, W. (2017). A Comparative Study of Feature Selection Methods for the Discriminative Analysis of Temporal Lobe Epilepsy. Frontiers in Neurology, 8, p. 633. doi: 10.3389/fneur.2017.00633.

Lange, C. and Yee, D. (2008). Progesterone and Breast Cancer. Women's Health, 4(2), pp.151-162. doi: 10.2217/17455057.4.2.151.

Larsen, S., Gao, Y. and Basse, P. (2014). NK Cells in the Tumor Microenvironment. Critical Reviews in Oncogenesis, 19(1-2), pp.91-105. doi: 10.1615/critrevoncog.2014011142.

Laywell, E., Kearns, S., Zheng, T., Chen, K., Deng, J., Chen, H., Roper, S. and Steindler, D. (2005). Neuron-to-astrocyte transition: Phenotypic fluidity and the formation of hybrid asterons in differentiating neurospheres. The Journal of Comparative Neurology, 493(3), pp.321-333. doi.org/10.1002/cne.20722.

Lee, B., Lee, B., Iyer, V., Sleckman, B., Shaffer, A., Ippolito, G., Tucker, H. and Dekker, J. (2017). Corrected and Republished from: BCL11A Is a Critical Component of a Transcriptional Network That Activates RAG Expression and V(D)J Recombination. Molecular and Cellular Biology, 38(1), pe00362-17. doi: 10.1128/mcb.00362-17.

Leinonen, R., Sugawara, H. and Shumway, M. (2010). The Sequence Read Archive. Nucleic Acids Research, 39(Database), pp.D19-D21. doi: 10.1093/nar/gkq1019.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. and Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems, 1(6), pp.417-425. doi: 10.1016/j.cels.2015.12.004.

Liddelow, S. and Barres, B. (2017). Reactive Astrocytes: Production, Function, and Therapeutic Potential. Immunity, 46(6), pp.957-967. doi.org/10.1016/j.immuni.2017.06.006.

Lim, E., Palmieri, C. and Tilley, W. (2016). Renewed interest in the progesterone receptor in breast cancer. British Journal of Cancer, 115(8), pp.909-911. doi: 10.1038/bjc.2016.303.

Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J., Li, B. and Liu, X. (2017). TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. Cancer Research, 77(21), pp.e108-e110. doi: 10.1158/0008-5472.can-17-0307.

Liu, Z., Zhang, X. and Zhang, S. (2014). Breast tumor subgroups reveal diverse clinical prognostic power. Scientific Reports, 4(1). doi.org/10.1038/srep04002.

Lonsdale, J., Thomas, J., Salvatore, … and Moore, H. (2013). The Genotype-Tissue Expression (GTEx) project. Nature Genetics, 45(6), pp.580-585. doi: 10.1038/ng.2653.

Louveau, A., Harris, T. and Kipnis, J. (2015). Revisiting the Mechanisms of CNS Immune Privilege. Trends in Immunology, 36(10), pp.569-577. doi: 10.1016/j.it.2015.08.006.

Luciano-Montalvo, C. and Meléndez, L. (2009). Cystatin B Associates with Signal Transducer and Activator of Transcription 1 in Monocyte-Derived and Placental Macrophages. Placenta, 30(5), pp.464-467.doi: 10.1016/j.placenta.2009.03.003.

Lugo-Villarino, G., Troegeler, A., Balboa, L., Lastrucci, C., Duval, C., Mercier, I., Bénard, A., Capilla, F., Al Saati, T., Poincloux, R., Kondova, I., Verreck, F., Cougoule, C., Maridonneau-Parini, I., Sasiain, M. and Neyrolles, O. (2018). The C-Type Lectin Receptor DC-SIGN Has an Anti-Inflammatory Role in Human M(IL-4) Macrophages in Response to Mycobacterium tuberculosis. Frontiers in Immunology, 9. doi: 10.3389/fimmu.2018.01123.

Lumachi, F., Brunello, A., Maruzzo, M., Basso, U. and Basso, S. (2013). Treatment of Estrogen Receptor-Positive Breast Cancer. Current Medicinal Chemistry, 20(5), pp.596-604. doi: 10.2174/092986713804999303.

Lun, A., Bach, K. and Marioni, J. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biology, 17(1), p.75. doi.org/10.1186/s13059-016-0947-7.

M

Mabuchi, S., Yokoi, E., Komura, N. and Kimura, T. (2018). Myeloid-derived suppressor cells and their role in gynecological malignancies. Tumor Biology, 40(7), p.101042831877648. doi: 10.1177/1010428318776485.

Mailman, M., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., Popova, N., Pretel, S., Ziyabari, L., Lee, M., Shao, Y., Wang, Z., Sirotkin, K., Ward, M., Kholodov, M., Zbicz, K., Beck, J., Kimelman, M., Shevelev, S., Preuss, D., Yaschenko, E., Graeff, A., Ostell, J. and Sherry, S. (2007). The NCBI dbGaP database of genotypes and phenotypes. Nature Genetics, 39(10), pp.1181-1186. doi: 10.1038/ng1007-1181.

Maimela, N., Liu, S. and Zhang, Y. (2019). Fates of CD8+ T cells in Tumor Microenvironment. Computational and Structural Biotechnology Journal, 17, pp.1-13. doi: 10.1016/j.csbj.2018.11.004.

Makhoul, I., Atiq, M., Alwbari, A. and Kieber-Emmons, T. (2018). Breast Cancer Immunotherapy: An Update. Breast Cancer: Basic and Clinical Research, 12, p.117822341877480. doi: 10.1177/1178223418774802.

Makin, S. (2018). The amyloid hypothesis on trial. Nature, 559(7715), pp.S4-S7. doi: 10.1038/d41586-018-05719-4.

Makki, J. (2015). Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance. Clinical Medicine Insights: Pathology, 8, p.CPath.S31563. doi: 10.4137/CPath.S31563.

Man, Y., Stojadinovic, A., Mason, J., Avital, I., Bilchik, A., Bruecher, B., Protic, M., Nissan, A., Izadjoo, M., Zhang, X. and Jewett, A. (2013). Tumor-Infiltrating Immune Cells Promoting Tumor Invasion and Metastasis: Existing Theories. Journal of Cancer, 4(1), pp.84-95. doi:10.7150/jca.5482.

Marshall, A., Du, Q., Draves, K., Shikishima, Y., HayGlass, K. and Clark, E. (2002). FDC-SP, a Novel Secreted Protein Expressed by Follicular Dendritic Cells. The Journal of Immunology, 169(5), pp.2381-2389. doi.org/10.4049/jimmunol.169.5.2381.

Martinez, F. and Gordon, S. (2014). The M1 and M2 paradigm of macrophage activation: time for reassessment. F1000Prime Reports, 6. doi: 10.12703/P6-13.

Mellman, I. (2013). Dendritic Cells: Master Regulators of the Immune Response. Cancer Immunology Research, 1(3), pp.145-149. doi: 10.1158/2326-6066.CIR-13-0102.

Mertens, C., Akam, E., Rehwald, C., Brüne, B., Tomat, E. and Jung, M. (2016). Intracellular Iron Chelation Modulates the Macrophage Iron Phenotype with Consequences on Tumor Progression. PLOS ONE, 11(11), p.e0166164. doi: 10.1371/journal.pone.0166164.

Mi, H., Muruganujan, A., Ebert, D., Huang, X. and Thomas, P. (2018). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Research, 47(D1), pp.D419-D426. doi.org/10.1093/nar/gky1038.

Milioli, H., Tishchenko, I., Riveros, C., Berretta, R. and Moscato, P. (2017). Basal-like breast cancer: molecular profiles, clinical features and survival outcomes. BMC Medical Genomics, 10(1). doi: 10.1186/s12920-017-0250-9.

Mitri, Z., Constantine, T. and O'Regan, R. (2012). The HER2 Receptor in Breast Cancer: Pathophysiology, Clinical Use, and New Advances in Therapy. Chemotherapy Research and Practice, 2012, pp.1-7. doi: 10.1155/2012/743193.

Montecino-Rodriguez, E., Berent-Maoz, B. and Dorshkind, K. (2013). Causes, consequences, and reversal of immune system aging. Journal of Clinical Investigation, 123(3), pp.958-965. doi: 10.1172/jci64096.

Mukhin, V., Pavlov, K. and Klimenko, V. (2017). Mechanisms of Neuron Loss in Alzheimer's Disease. Neuroscience and Behavioral Physiology, 47(5), pp.508-516. doi.org/10.1007/s11055-017-0427-x.

N

National Cancer Institute. (2019). Cancer Statistics. [online] Available at: https://www.cancer.gov/about-cancer/understanding/statistics [Accessed 5 Sep. 2019].

Nayak, D., Roth, T. and McGavern, D. (2014). Microglia Development and Function. Annual Review of Immunology, 32(1), pp.367-402. doi: 10.1146/annurev-immunol-032713-120240.

Ncbi.nlm.nih.gov. (2019). HLA-DOB major histocompatibility complex, class II, DO beta [Homo sapiens (human)] - Gene - NCBI. [online] Available at: https://www.ncbi.nlm.nih.gov/gene/3112 [Accessed 21 Sep. 2019].

Newman, A., Steen, C., Liu, C., Gentles, A., Chaudhuri, A., Scherer, F., Khodadoust, M., Esfahani, M., Luca, B., Steiner, D., Diehn, M. and Alizadeh, A. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature Biotechnology, 37(7), pp.773-782.

O

Omilusik, K. and Goldrath, A. (2017). The origins of memory T cells. Nature, 552(7685), pp.337-339. doi: 10.1038/d41586-017-08280-8.

Omim.org. (2019). OMIM Entry - * 600108 - MATRIX METALLOPROTEINASE 13; MMP13. [online] Available at: https://omim.org/entry/600108 [Accessed 21 Sep. 2019].

Omim.org. (2019). OMIM Entry - * 606557 - BAF CHROMATIN REMODELING COMPLEX SUBUNIT BCL11A; BCL11A. [online] Available at: https://omim.org/entry/606557 [Accessed 21 Sep. 2019].

Omim.org. (2019). OMIM Entry - * 607241 - CHROMOSOME 4 OPEN READING FRAME 7; C4ORF7. [online] Available at: https://omim.org/entry/607241 [Accessed 21 Sep. 2019].

P

Palmer, A. and Ousman, S. (2018). Astrocytes and Aging. Frontiers in Aging Neuroscience, 10, p.337. doi.org/10.3389/fnagi.2018.00337.

Papalexi, E. and Satija, R. (2017). Single-cell RNA sequencing to explore immune cell heterogeneity. Nature Reviews Immunology, 18(1), pp.35-45. doi: 10.1038/nri.2017.76.

Park, I., Hwang, S., Song, I., Heo, S., Kim, Y., Bang, W., Park, H., Lee, M., Gong, G. and Lee, H. (2017). Expression of the MHC class II in triple-negative breast cancer is associated with tumor-infiltrating lymphocytes and interferon signaling. PLOS ONE, 12(8), p.e0182786. doi: 10.1371/journal.pone.0182786.

Pau G. and Reeder J. HTSeqGenie: A NGS analysis pipeline. R package version 3.16.0 (2014)

Pavlichin, D. and Weissman, T. (2016). Chained Kullback-Leibler divergences. International Symposium on Information Theory (ISIT). pp. 580–584. doi: 10.1109/isit.2016.7541365.

Pawelec, G. (2017). Immunosenescence and cancer. Biogerontology, 18(4), pp.717-721. doi: 10.1007/s10522-017-9682-z.

Pfenninger, C., Roschupkina, T., Hertwig, F., Kottwitz, D., Englund, E., Bengzon, J., Jacobsen, S. and Nuber, U. (2007). CD133 Is Not Present on Neurogenic Astrocytes in the Adult Subventricular Zone, but on Embryonic Neural Stem Cells, Ependymal Cells, and Glioblastoma Cells. Cancer Research, 67(12), pp.5727-5736. doi:10.1158/0008-5472.can-07-0183.

Pierce, A., Bullain, S. and Kawas, C. (2017). Late-Onset Alzheimer Disease. Neurologic Clinics, 35(2), pp.283-293. doi: 10.1016/j.ncl.2017.01.006.

Poh, A. and Ernst, M. (2018). Targeting Macrophages in Cancer: From Bench to Bedside. Frontiers in Oncology, 8, p.49. doi: 10.3389/fonc.2018.00049.

Provinciali, M., Pierpaoli, E., Malavolta, M., Donnini, A., Smorlesi, A. and Gatti, C. (2017). Breast Cancer and Immunosenescence. Handbook of Immunosenescence, pp.1-31.

Q

Qian, L. and Flood, P. (2008). Microglial cells and Parkinson's disease. Immunologic Research, 41(3), pp.155-164. doi: 10.1007/s12026-008-8018-0.

Qiao, W., Quon, G., Csaszar, E., Yu, M., Morris, Q. and Zandstra, P. (2012). PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and

Developmental Conditions. PLoS Computational Biology, 8(12), p.e1002838. doi: 10.1371/journal.pcbi.1002838.

Qiu, S., Waaijer, S., Zwager, M., de Vries, E., van der Vegt, B. and Schröder, C. (2018). Tumor-associated macrophages in breast cancer: Innocent bystander or important player? Cancer Treatment Reviews, 70, pp.178-189. doi.org/10.1016/j.ctrv.2018.08.010.

R

Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. and Gfeller, D. (2017). Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. eLife, 6, p.e26476. doi: 10.7554/eLife.26476.

Räisänen, S., Alatalo, S., Ylipahkala, H., Halleen, J., Cassady, A., Hume, D. and Väänänen, H. (2005). Macrophages overexpressing tartrate-resistant acid phosphatase show altered profile of free radical production and enhanced capacity of bacterial killing. Biochemical and Biophysical Research Communications, 331(1), pp.120-126. doi.org/10.1016/j.bbrc.2005.03.133.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at: https://www.R-project.org.

Reizis, B. (2019). Plasmacytoid Dendritic Cells: Development, Regulation, and Function. Immunity, 50(1), pp.37-50. doi: 10.1016/j.immuni.2018.12.027.

Ritchie, M., Phipson, B., Wu, D., Hu, Y., Law, C., Shi, W. and Smyth, G. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7), pp.e47-e47. doi.org/10.1093/nar/gkv007.

Roche, J. (2018). The Epithelial-to-Mesenchymal Transition in Cancer. Cancers, 10(2), p.52.doi: 10.3390/cancers10020052.

S

Seachrist, D., Ingles, N., Hannigan, M., Licatalosi, D. and Keri, R. (2018). BCL11A is necessary for the expression of extracellular matrix genes and metastatic progression of triple-negative breast cancer. Tumor Biology. 78(13), Abstract nr 32. doi.10.1158/1538-7445.am2018-32.

See, P., Lum, J., Chen, J. and Ginhoux, F. (2018). A Single-Cell Sequencing Guide for Immunologists. Frontiers in Immunology, 9, p.2425. doi: 10.3389/fimmu.2018.02425.

Sekula, M., Gaskins, J. and Datta, S. (2019). Detection of differentially expressed genes in discrete single-cell RNA sequencing data using a hurdle model with correlated random effects. Biometrics. doi.org/10.1111/biom.13074.

Schachter, A. S., and Davis, K. L. (2000). Alzheimer's disease. Dialogues in clinical neuroscience, 2(2), pp.91–100.

Schelker, M., Feau, S., Du, J., Ranu, N., Klipp, E., MacBeath, G., Schoeberl, B. and Raue, A. (2017). Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nature Communications, 8(1), p.2032. doi: 10.1038/s41467-017-02289-3.

Schiweck, J., Eickholt, B. and Murk, K. (2018). Important Shapeshifter: Mechanisms Allowing Astrocytes to Respond to the Changing Nervous System During Development, Injury and Disease. Frontiers in Cellular Neuroscience, 12, p.261. doi.org/10.3389/fncel.2018.00261.

Sharma, P. and Allison, J. (2015). Immune Checkpoint Targeting in Cancer Therapy: Toward Combination Strategies with Curative Potential. Cell, 161(2), pp.205-214.

Shechter, R., Miller, O., Yovel, G., Rosenzweig, N., London, A., Ruckh, J., Kim, K., Klein, E., Kalchenko, V., Bendel, P., Lira, S., Jung, S. and Schwartz, M. (2013). Recruitment of Beneficial M2 Macrophages to Injured Spinal Cord Is Orchestrated by Remote Brain Choroid Plexus. Immunity, 38(3), pp.555-569. doi: 10.1016/j.immuni.2013.02.012.

Shults, C. (2006). Lewy bodies. Proceedings of the National Academy of Sciences, 103(6), pp.1661-1668. doi: 10.1073/pnas.0509567103.

Smolders, J., Heutinck, K., Fransen, N., Remmerswaal, E., Hombrink, P., ten Berge, I., van Lier, R., Huitinga, I. and Hamann, J. (2018). Tissue-resident memory T cells populate the human brain. Nature Communications, 9(1), p.4593. doi.org/10.1038/s41467-018-07053-9.

Smyth, M., Dunn, G. and Schreiber, R. (2006). Cancer Immunosurveillance and Immunoediting: The Roles of Immunity in Suppressing Tumor Development and Shaping Tumor Immunogenicity. Advances in Immunology, pp.1-50. doi.org/10.1016/S0065-2776(06)90001-7.

Sorlie, T., Perou, C., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M., van de Rijn, M., Jeffrey, S., Thorsen, T., Quist, H., Matese, J., Brown, P., Botstein, D., Lonning, P. and Borresen-Dale, A. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences, 98(19), pp.10869-10874. doi: 10.1073/pnas.191367098.

Spaethling, J., Na, Y., Lee, J., Ulyanova, A., Baltuch, G., Bell, T., Brem, S., Chen, H., Dueck, H., Fisher, S., Garcia, M., Khaladkar, M., Kung, D., Lucas, T., O'Rourke, D., Stefanik, D., Wang, J., Wolf, J., Bartfai, T., Grady, M., Sul, J., Kim, J. and Eberwine, J. (2017). Primary Cell Culture of Live Neurosurgically Resected Aged Adult Human Brain Cells and Single Cell Transcriptomics. Cell Reports, 18(3), pp.791-803. doi: 10.1016/j.celrep.2016.12.066.

Srinivasan, K., Friedman, B., Etxeberria, A., Huntley, M., van der Brug, M., Foreman, O., Paw, J., Modrusan, Z., Beach, T., Serrano, G. and Hansen, D. (2019). Alzheimer's patient brain myeloid cells exhibit enhanced aging and unique transcriptional activation. doi:10.1101/610345.

Streets, A. and Huang, Y. (2014). How deep is enough in single-cell RNA-seq? Nature Biotechnology, 32(10), pp.1005-1006. doi: 10.1038/nbt.3039.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W., Hao, Y., Stoeckius, M., Smibert, P. and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell, 177(7), pp.1888-1902.e21. doi.org/10.1016/j.cell.2019.05.031.

Stubbington, M., Rozenblatt-Rosen, O., Regev, A. and Teichmann, S. (2017). Single-cell transcriptomics to explore the immune system in health and disease. Science, 358(6359), pp.58-63. doi.10.1126/science.aan6828.

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E. and Mesirov, J. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43), pp.15545-15550. doi.org/10.1073/pnas.0506580102.

Sun, Y., Zhao, Z., Yang, Z., Xu, F., Lu, H., Zhu, Z., Shi, W., Jiang, J., Yao, P. and Zhu, H. (2017). Risk Factors and Preventions of Breast Cancer. International Journal of Biological Sciences, 13(11), pp.1387-1397. doi: 10.7150/ijbs.21635.

Svensson, V., Vento-Tormo, R. and Teichmann, S. (2018). Exponential scaling of single-cell RNA-seq in the past decade. Nature Protocols, 13(4), pp.599-604. doi: 10.1038/nprot.2017.

T

Andrews, T. and Hemberg, M. (2018). Identifying cell populations with scRNASeq. Molecular Aspects of Medicine, 59, pp.114-122. doi.org/10.1016/j.mam.2017.07.002.

Tanabe, S. and Yamashita, T. (2018). The role of immune cells in brain development and neurodevelopmental diseases. International Immunology, 30(10), pp.437-444. doi: 10.1093/intimm/dxy041.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B., Siddiqui, A., Lao, K. and Surani, M. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods, 6(5), pp.377-382. doi: 10.1038/nmeth.1315.

Tang, R., Ojeda, M., Pouillart, P., Scholl, S., Beuvon, F. and Mosseri, V. (1992). M-CSF (monocyte colony stimulating factor) and M-CSF receptor expression by breast tumour cells: M-CSF mediated recruitment of tumour infiltrating monocytes? Journal of Cellular Biochemistry, 50(4), pp.350-356. doi.org/10.1002/jcb.240500403.

Tasic, B., Menon, V., Nguyen, T., Kim, T., Jarsky, T., Yao, Z., Levi, B., Gray, L., Sorensen, S., Dolbeare, T., Bertagnolli, D., Goldy, J., Shapovalova, N., Parry, S., Lee, C., Smith, K., Bernard, A., Madisen, L., Sunkin, S., Hawrylycz, M., Koch, C. and Zeng, H. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nature Neuroscience, 19(2), pp.335-346.

Tham, C., Lin, F., Rao, T., Yu, N. and Webb, M. (2003). Microglial activation state and lysophospholipid acid receptor expression. International Journal of Developmental Neuroscience, 21(8), pp.431-443. doi.org/10.1016/j.ijdevneu.2003.09.003.

The GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nature Genetics, 45(6), pp.580-585. doi: 10.1038/ng.2653.

Therneau, T. (2015). A Package for Survival Analysis in S_. version 2.38. Available online at: https://CRAN.R-project.org/package=survival.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences, 99(10), pp.6567-6572. doi: 10.1073/pnas.082099299.

Tomczak, K., Czerwińska, P. and Wiznerowicz, M. (2015). Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Współczesna Onkologia, 1A, pp.68-77. doi: 10.5114/wo.2014.47136.

Tong, C., Wu, M., Cho, W. and To, K. (2018). Recent Advances in the Treatment of Breast Cancer. Frontiers in Oncology, 8. doi: 10.3389/fonc.2018.00227.

Torre-Minguela, C., Barberà-Cremades, M., Gómez, A., Martín-Sánchez, F. and Pelegrín, P. (2016). Macrophage activation and polarization modify P2X7 receptor secretome influencing the inflammatory process. Scientific Reports, 6(1), p.22586. doi.org/10.1038/srep22586.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Research, 25(10), pp.1491-1498. doi:10.1101/gr.190595.115.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N., Livak, K., Mikkelsen, T. and Rinn, J. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology, 32(4), pp.381-386. doi.org/10.1038/nbt.2859.

U

United Nations, Department of Economic and Social Affairs, Population Division (2017). World Population Ageing 2017 - Highlights (ST/ESA/SER.A/397).

Ushach, I. and Zlotnik, A. (2016). Biological role of granulocyte macrophage colony-stimulating factor (GM-CSF) and macrophage colony-stimulating factor (M-CSF) on cells of the myeloid lineage. Journal of Leukocyte Biology, 100(3), pp.481-489. doi: 10.1189/jlb.3ru0316-144r.

V

van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research, 9, pp.2579-2605.

Vasile, F., Dossi, E. and Rouach, N. (2017). Human astrocytes: structure and functions in the healthy brain. Brain Structure and Function, 222(5), pp.2017-2029. doi: 10.1007/s00429-017-1383-5.

Verdot, L., Lalmanach, G., Vercruysse, V., Hartmann, S., Lucius, R., Hoebeke, J., Gauthier, F. and Vray, B. (1996). Cystatins Up-regulate Nitric Oxide Release from Interferon-γ- activated Mouse Peritoneal Macrophages. Journal of Biological Chemistry, 271(45), pp.28077-28081. doi: 10.1074/jbc.271.45.28077.

Vies, S. M. (2016). Protein kinases orchestrate early pathological events in Alzheimer's disease.

Visanji, N., Brooks, P., Hazrati, L. and Lang, A. (2013). The prion hypothesis in Parkinson's disease: Braak to the future. Acta Neuropathologica Communications, 1(1), p.2. doi: 10.1186/2051-5960-1-2.

Vivier, E., Tomasello, E., Baratin, M., Walzer, T. and Ugolini, S. (2008). Functions of natural killer cells. Nature Immunology, 9(5), pp.503-510. doi: 10.1038/ni1582.

von Bartheld, C., Bahney, J. and Herculano-Houzel, S. (2016). The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. Journal of Comparative Neurology, 524(18), pp.3865-3895. doi: 10.1002/cne.24040.

W

Wagner, A., Regev, A. and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. Nature Biotechnology, 34(11), pp.1145-1160. doi.org/10.1038/nbt.3711.

Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C. and Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. Frontiers in Genetics, 4, p.270. doi: 10.3389/fgene.2013.00270.

Wang, C., Zhou, L., Li, S., Wei, J., Wang, W., Zhou, T., Liao, S., Weng, D., Deng, D., Weng, Y., Wang, S. and Ma, D. (2010). C4orf7 contributes to ovarian cancer metastasis by promoting cancer cell migration and invasion. Oncology Reports. 24(4), pp. 933-939.doi.org/10.3892/or_00000939.

Wang, J., Huang, M., Torre, E., Dueck, H., Shaffer, S., Murray, J., Raj, A., Li, M. and Zhang, N. (2018). Gene expression distribution deconvolution in single-cell RNA sequencing. Proceedings of the National Academy of Sciences, 115(28), pp.E6437-E6446. doi.org/10.1073/pnas.1721085115.

Wang, M., Zhao, J., Zhang, L., Wei, F., Lian, Y., Wu, Y., Gong, Z., Zhang, S., Zhou, J., Cao, K., Li, X., Xiong, W., Li, G., Zeng, Z. and Guo, C. (2017). Role of tumor microenvironment in tumorigenesis. Journal of Cancer, 8(5), pp.761-773. doi: 10.7150/jca.17648.

Wang, X., Park, J., Susztak, K., Zhang, N. and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature Communications, 10(1), p.380. doi.org/10.1038/s41467-018-08023-x.

Wang, Z., Gerstein, M. and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1), pp.57-63. doi: 10.1038/nrg2484.

Ward, R., Zucca, F., Duyn, J., Crichton, R. and Zecca, L. (2014). The role of iron in brain ageing and neurodegenerative disorders. The Lancet Neurology, 13(10), pp.1045-1060. doi: 10.1016/S1474-4422(14)70117-6.

Weagel E., Smith C., Liu P., Robison R., and O'Neill K. (2015). Macrophage Polarization and Its Role in Cancer. Journal of Clinical & Cellular Immunology, 06(04), p.338. doi:10.4172/2155-9899.1000338.

Weinstein, J., Collisson, E., Mills, G., Shaw, K., Ozenberger, B., Ellrott, K., Shmulevich, I., Sander, C. and Stuart, J. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics, 45(10), pp.1113-1120. doi.org/10.1038/ng.2764.

Wesolowski, R. and Ramaswamy, B. (2018). Gene expression profiling: changing face of breast cancer classification and management. Gene expression, 15(3), pp.105-115.

WHO, US National Institute of Aging (2011). Global Health and Aging (NIH Publication no. 11-7737).

Wu, J. and Lanier, L. (2003). Natural Killer Cells and Cancer. Advances in Cancer Research, pp.127-156. doi.org/10.1016/s0065-230x(03)90004-2.

Wu, T., Reeder, J., Lawrence, M., Becker, G. and Brauer, M. (2016). GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. Methods in Molecular Biology, pp.283-334. doi: 10.1007/978-1-4939-3578-9_15.

Wyss-Coray, T. (2016). Ageing, neurodegeneration and brain rejuvenation. Nature, 539(7628), pp.180-186. doi:10.1038/nature20411.

Y

Yamazaki, Y. and Kanekiyo, T. (2017). Blood-Brain Barrier Dysfunction and the Pathogenesis of Alzheimer's Disease. International Journal of Molecular Sciences, 18(9), p.1965. doi: 10.3390/ijms18091965.

Yankner, B., Duffy, L. and Kirschner, D. (1990). Neurotrophic and neurotoxic effects of amyloid beta protein: reversal by tachykinin neuropeptides. Science, 250(4978), pp.279-282. doi: 10.1126/science.221853.

Yankner, B., Lu, T. and Loerch, P. (2008). The Aging Brain. Annual Review of Pathology: Mechanisms of Disease, 3(1), pp.41-66. doi.org/10.1146/annurev.pathmechdis.2.010506.092044.

Yau, Y., Zeighami, Y., Baker, T., Larcher, K., Vainik, U., Dadar, M., Fonov, V., Hagmann, P., Griffa, A., Mišić, B., Collins, D. and Dagher, A. (2018). Network connectivity determines cortical thinning in early Parkinson's disease progression. Nature Communications, 9(1). doi.org/10.1038/s41467-017-02416-0.

Yersal, O. and Barutca, S. (2014). Biological subtypes of breast cancer: Prognostic and therapeutic implications. World Journal of Clinical Oncology, 5(3), p.412. doi: 10.5306/wjco.v5.i3.412.

Yu, Y., Wang, J., Khaled, W., Burke, S., Li, P., Chen, X., Yang, W., Jenkins, N., Copeland, N., Zhang, S. and Liu, P. (2012). Bcl11a is essential for lymphoid development and negatively regulates p53. The Journal of Experimental Medicine, 209(13), pp.2467-2483. doi: 10.1084/jem.20121846.

Z

Zhang, F., Wang, H., Wang, X., Jiang, G., Liu, H., Zhang, G., Wang, H., Fang, R., Bu, X., Cai, S. and Du, J. (2016). TGF-β induces M2-like macrophage polarization via SNAIL-mediated suppression of a pro-inflammatory phenotype. Oncotarget, 7(32), pp. 52294–52306. doi: 10.18632/oncotarget.10561.

Zhang, Y., Sloan, S., Clarke, L., Caneda, C., Plaza, C., Blumenthal, P., Vogel, H., Steinberg, G., Edwards, M., Li, G., Duncan, J., Cheshier, S., Shuer, L., Chang, E., Grant, G., Gephart, M. and Barres, B. (2016). Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. Neuron, 89(1), pp.37-53. doi: 10.1016/j.neuron.2015.11.013.

Zheng, H. and Koo, E. (2006). The amyloid precursor protein: beyond amyloid. Molecular Neurodegeneration, 1(1), p.5. doi: 10.1186/1750-1326-1-5.

Zinger, A., Cho, W. and Ben-Yehuda, A. (2017). Cancer and Aging - the Inflammatory Connection. Aging and Disease, 8(5), p.611. doi: 10.14336/ad.2016.1230.

10.Supplementary figures

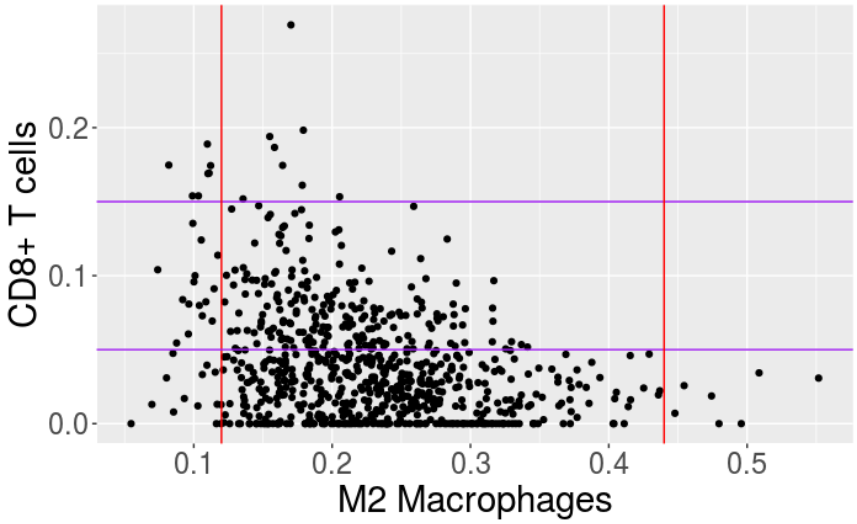

Figure 10.1: Selection of breast tumour bulk RNA-seq samples to perform differential expressional analysis, as described in section 6.2.8. Red highlights cut-offs for the proportion of M2 macrophages; Purple highlights cut-offs for the proportion of CD8+ T cells.

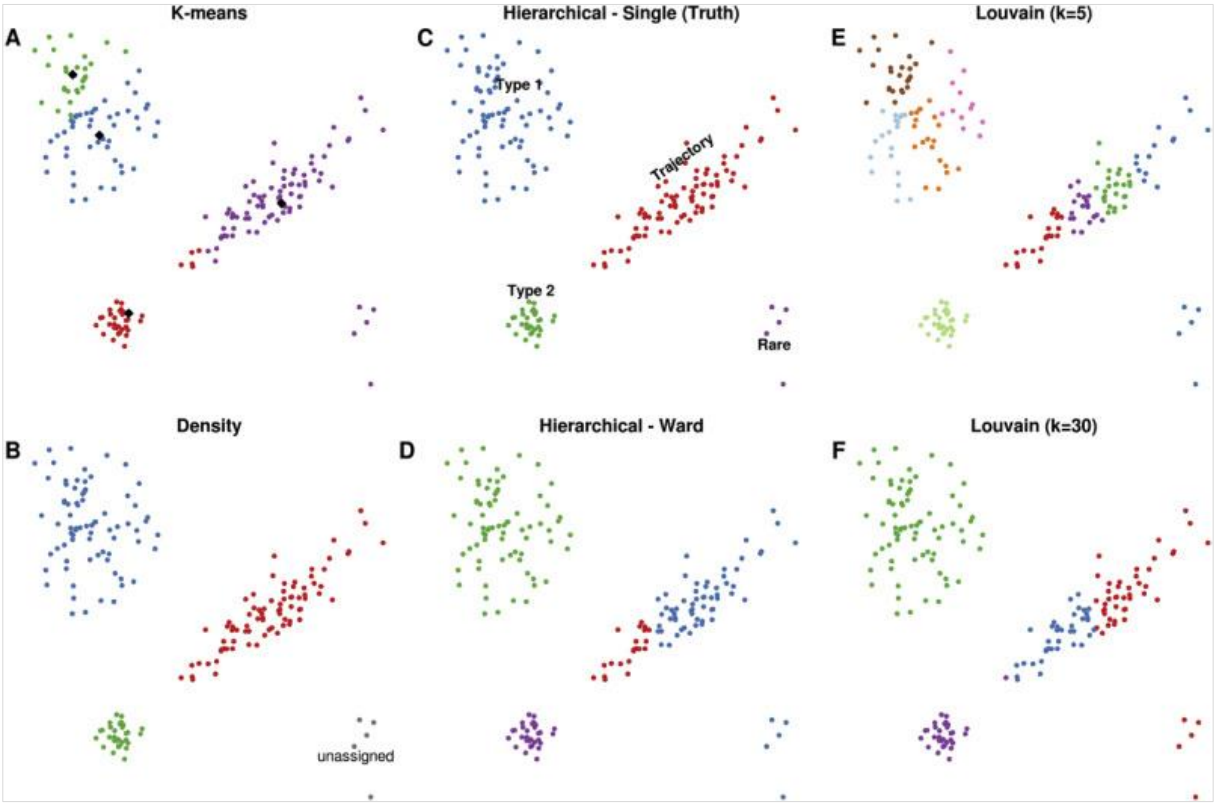

Figure 10.2: t-SNEs of different synthetic scRNA-seq data structures obtained with different clustering methods. ***Type 1*** *– noisy population;* ***Type 2*** *- developmental trajectory;* ***Rare*** *- rare cell population. From Tallulah and Hemberg, 2018.*

Table 10.1: Results from the GO enrichment analysis using marker genes of reactive astrocytes from the Spaethling dataset. This analysis was performed using the PANTHER Classification System (Mi et al. 2018) with GO annotations.

GO biological process	FDR
neuroblast differentiation (GO:0014016)	4.95E-02
astrocyte activation (GO:0048143)	1.70E-02
glial cell activation (GO:0061900)	9.13E-03
astrocyte development (GO:0014002)	9.71E-03
neuroinflammatory response (GO:0150076)	1.06E-02
glial cell migration (GO:0008347)	1.61E-02
positive regulation of neural precursor cell proliferation (GO:2000179)	4.53E-03
neuron recognition (GO:0008038)	2.67E-02
astrocyte differentiation (GO:0048708)	2.90E-02
glial cell development (GO:0021782)	8.65E-04
regulation of neural precursor cell proliferation (GO:2000177)	2.26E-02
gliogenesis (GO:0042063)	9.93E-06
glial cell differentiation (GO:0010001)	1.67E-03
regulation of neurological system process (GO:0031644)	1.67E-02
negative regulation of neuron differentiation (GO:0045665)	3.46E-04
negative regulation of nervous system development (GO:0051961)	1.58E-04
negative regulation of neurogenesis (GO:0050768)	3.97E-04
negative regulation of cell development (GO:0010721)	7.03E-04
axon development (GO:0061564)	7.10E-04
positive regulation of neurogenesis (GO:0050769)	2.18E-03

Table 10.2: Results from the GO enrichment analysis using marker genes of resting astrocytes from the Darmanis dataset. This analysis was performed using the PANTHER Classification System (Mi et al. 2018) with GO annotations.

GO biological process complete	FDR
retinal rod cell differentiation (GO:0060221)	2.41E-02
dipeptide transport (GO:0042938)	2.39E-02
dipeptide transmembrane transport (GO:0035442)	2.37E-02
seminal vesicle development (GO:0061107)	2.36E-02
D-aspartate import across plasma membrane (GO:0070779)	3.19E-02
D-aspartate transport (GO:0070777)	3.17E-02
L-glutamate import across plasma membrane (GO:0098712)	2.62E-03
L-glutamate import (GO:0051938)	3.14E-03
glutamate biosynthetic process (GO:0006537)	4.99E-02
L-glutamate transmembrane transport (GO:0015813)	1.24E-03
glial cell fate commitment (GO:0021781)	9.43E-03
amino acid import across plasma membrane (GO:0089718)	9.34E-03
cell communication by electrical coupling involved in cardiac conduction (GO:0086064)	1.09E-02
cellular sodium ion homeostasis (GO:0006883)	1.39E-02
neurotransmitter uptake (GO:0001504)	2.05E-03
cell communication by electrical coupling (GO:0010644)	1.56E-02
negative regulation of adenylate cyclase activity (GO:0007194)	1.77E-02
amino acid import (GO:0043090)	1.76E-02
nitric oxide mediated signal transduction (GO:0007263)	2.00E-02
positive regulation of chondrocyte differentiation (GO:0032332)	2.23E-02